CALCULUS ILLUSTRATED

VOLUME 1: PRECALCULUS

PETER SAVELIEV

To the student

Mathematics is a science. Just as the rest of the scientists, mathematicians are trying to understand how the Universe operates and discover its laws. When successful, they write these laws as short statements called "theorems". In order to present these laws conclusively and precisely, a dictionary of the new concepts is also developed; its entries are called "definitions". These two make up the most important part of any mathematics book.

This is how definitions, theorems, and some other items are used as building blocks of the scientific theory we present in this text.

Every new concept is introduced with utmost specificity.

Definition 0.0.1: square root

Suppose a is a positive number. Then the *square root* of a is a positive number x, such that $x^2 = a$.

The term being introduced is given in *italics*. The definitions are then constantly referred to throughout the text.

New symbolism may also be introduced.

Square root

$$\sqrt{a}$$

Consequently, the notation is freely used throughout the text.

We may consider a specific instance of a new concept either before or after it is explicitly defined.

Example 0.0.2: length of diagonal

What is the length of the diagonal of a 1×1 square? The square is made of two right triangles and the diagonal is their shared hypotenuse. Let's call it a. Then, by the *Pythagorean Theorem*, the square of a is $1^2 + 1^2 = 2$. Consequently, we have:

$$a^2 = 2\,.$$

We immediately see the need for the square root! The length is, therefore, $a = \sqrt{2}$.

You can skip some of the examples without violating the flow of ideas, at your own risk.

All new material is followed by a few little tasks, or questions, like this.

Exercise 0.0.3

Find the height of an equilateral triangle the length of the side of which is 1.

The exercises are to be attempted (or at least considered) immediately.

Most of the in-text exercises are not elaborate. They aren't, however, entirely routine as they require understanding of, at least, the concepts that have just been introduced. Additional exercise *sets* are placed in the appendix as well as at the book's website: calculus123.com. Do not start your study with the exercises! Keep in mind that the exercises are meant to test – indirectly and imperfectly – how well the *concepts* have been learned.

There are sometimes words of caution about common mistakes made by the students.

Warning!

In spite of the fact that $(-1)^2 = 1$, there is only one square root of 1, $\sqrt{1} = 1$.

The most important facts about the new concepts are put forward in the following manner.

Theorem 0.0.4: Product of Roots

For any two positive numbers a and b, we have the following identity:

$$\sqrt{a} \cdot \sqrt{b} = \sqrt{a \cdot b}$$

The theorems are constantly referred to throughout the text.

As you can see, theorems may contain formulas; a theorem supplies limitations on the applicability of the formula it contains. Furthermore, every formula is a part of a theorem, and using the former without knowing the latter is perilous.

There is no need to memorize definitions or theorems (and formulas), initially. With enough time spent with the material, the main ones will eventually become familiar as they continue to reappear in the text. Watch for words "important", "crucial", etc. Those new concepts that do not reappear in this text are likely to be seen in the next mathematics book that you read. You need to, however, be aware of all of the definitions and theorems and be able to find the right one when necessary.

Often, but not always, a theorem is followed by a thorough argument as a justification.

Proof.

Suppose $A = \sqrt{a}$ and $B = \sqrt{b}$. Then, according to the definition, we have the following:

$$a = A^2 \quad \text{and} \quad b = B^2 \ .$$

Therefore, we have:

$$a \cdot b = A^2 \cdot B^2 = A \cdot A \cdot B \cdot B = (A \cdot B) \cdot (A \cdot B) = (AB)^2 \ .$$

Hence, $\sqrt{ab} = A \cdot B$, again according to the definition.

Some proofs can be skipped at first reading.

Its highly detailed exposition makes the book a good choice for *self-study*. If this is your case, these are my suggestions.

While reading the book, try to make sure that you understand new concepts and ideas. Keep in mind, however, that some are more important that others; they are marked accordingly. Come back (or jump forward) as needed. Contemplate. Find other sources if necessary. You should not turn to the exercise sets until you have become comfortable with the material.

What to do about exercises when solutions aren't provided? First, use the examples. Many of them contain a problem – with a solution. Try to solve the problem – before or after reading the solution. You can also find exercises online or make up your own problems and solve them!

I strongly suggest that your solution should be thoroughly *written*. You should write in complete sentences, including all the algebra. For example, you should appreciate the difference between these two:

Wrong: $\boxed{\begin{array}{l}1+1\\2\end{array}}$ Right: $\boxed{\begin{array}{l}1+1\\=2\end{array}}$

The latter reads "one added to one is two", while the former cannot be read. You should also justify all your steps and conclusions, including all the algebra. For example, you should appreciate the difference between these two:

Wrong: $\boxed{\begin{array}{l} 2x = 4 \\ x = 2 \end{array}}$ Right: $\boxed{\begin{array}{l} 2x = 4;\ \text{therefore,} \\ x = 2\,. \end{array}}$

The standards of thoroughness are provided by the examples in the book.

Next, your solution should be thoroughly *read*. This is the time for self-criticism: Look for errors and weak spots. It should be re-read and then rewritten. Once you are convinced that the solution is correct and the presentation is solid, you may show it to a knowledgeable person for a once-over.

Next, you may turn to modeling projects. Spreadsheets (Microsoft Excel or similar) are chosen to be used for graphing and modeling. One can achieve as good results with packages specifically designed for these purposes, but spreadsheets provide a tool with a wider scope of applications. Programming is another option.

Good luck!

To the teacher

The bulk of the material in the book comes from my lecture notes.

There is little emphasis on closed-form computations and algebraic manipulations. I do think that a person who has never integrated by hand (or differentiated, or applied the quadratic formula, etc.) cannot possibly understand integration (or differentiation, or quadratic functions, etc.). However, a large proportion of time and effort can and should be directed toward:

- understanding of the concepts and
- modeling in realistic settings.

The challenge of this approach is that it requires more abstraction rather than less.

Visualization is the main tool used to deal with this challenge. Illustrations are provided for every concept, big or small. The pictures that come out are sometimes very precise but sometimes serve as mere metaphors for the concepts they illustrate. The hope is that they will serve as visual "anchors" in addition to the words and formulas.

It is unlikely that a person who has never plotted the graph of a function by hand can understand graphs or functions. However, what if we want to plot more than just a few points in order to visualize curves, surfaces, vector fields, etc.? Spreadsheets were chosen over graphic calculators for visualization purposes because they represent the shortest step away from pen and paper. Indeed, the data is plotted in the simplest manner possible: one cell - one number - one point on the graph. For more advanced tasks such as modeling, spreadsheets were chosen over other software and programming options for their wide availability and, above all, their simplicity. Nine out of ten, the spreadsheet shown was initially created from scratch in front of the students who were later able to follow my footsteps and create their own.

About the tests. The book isn't designed to prepare the student for some preexisting exam; on the contrary, assignments should be based on what has been learned. The students' understanding of the concepts needs to be tested but, most of the time, this can be done only indirectly. Therefore, a certain share of routine, mechanical problems is inevitable. Nonetheless, no topic deserves more attention just because it's likely to be on the test.

If at all possible, don't make the students memorize formulas.

In the order of topics, the main difference from a typical calculus textbook is that sequences come before everything else. The reasons are the following:

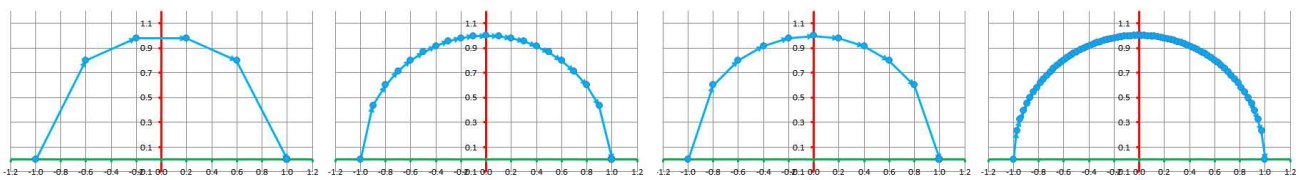
- Sequences are the simplest kind of functions.
- Limits of sequences are simpler than limits of general functions (including the ones at infinity).
- The sigma notation, the Riemann sums, and the Riemann integral make more sense to a student with a solid background in sequences.
- A quick transition from sequences to series often leads to confusion between the two.
- Sequences are needed for modeling, which should start as early as possible.

From the discrete to the continuous

It's no secret that a vast majority of calculus students will never use what they have learned. Poor career choices aside, a former calculus student is often unable to recognize the mathematics that is supposed to surround him. Why does this happen?

Calculus is the science of change. From the very beginning, its peculiar challenge has been to study and measure *continuous* change: curves and motion along curves. These curves and this motion are represented by *formulas*. Skillful manipulation of those formulas is what solves calculus problems. For over 300 years, this approach has been extremely successful in sciences and engineering. The successes are well-known: projectile motion, planetary motion, flow of liquids, heat transfer, wave propagation, etc. Teaching calculus follows this approach: An overwhelming majority of what the student does is manipulation of formulas on a piece of paper. But this means that all the problems the student faces were (or could have been) solved in the 18th or 19th centuries!

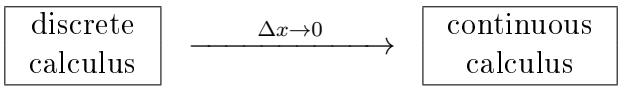
This isn't good enough anymore. What has changed since then? The computers have appeared, of course, and computers don't manipulate formulas. They don't help with solving – in the traditional sense of the word – those problems from the past centuries. Instead of *continuous*, computers excel at handling *incremental* processes, and instead of formulas they are great at managing discrete (digital) data. To utilize these advantages, scientists "discretize" the results of calculus and create algorithms that manipulate the digital data. The solutions are approximate but the applicability is unlimited. Since the 20th century, this approach has been extremely successful in sciences and engineering: aerodynamics (airplane and car design), sound and image processing, space exploration, structure of the atom and the universe, etc. The approach is also circuitous: Every concept in calculus *starts* – often implicitly – as a discrete approximation of a continuous phenomenon!



Calculus is the science of change, *both* incremental and continuous. The former part – the so-called discrete calculus – may be seen as the study of incremental phenomena and the quantities *indivisible* by their very nature: people, animals, and other organisms, moments of time, locations of space, particles, some commodities, digital images and other man-made data, etc. With the help of the calculus machinery called "limits", we invariably choose to transition to the continuous part of calculus, especially when we face continuous phenomena and the quantities *infinitely divisible* either by their nature or by assumption: time, space, mass, temperature, money, some commodities, etc. Calculus produces definitive results and absolute accuracy – but only for problems amenable to its methods! In the classroom, the problems are simplified until they become manageable; otherwise, we circle back to the discrete methods in search of approximations.

Within a typical calculus course, the student simply never gets to complete the "circle"! Later on, the graduate is likely to think of calculus only when he sees formulas and rarely when he sees numerical data.

In this book, every concept of calculus is first introduced in its discrete, "pre-limit", incarnation – elsewhere typically hidden inside proofs – and then used for modeling and applications well before its continuous counterpart emerges. The properties of the former are discovered first and then the matching properties of the latter are found by making the increment smaller and smaller, at the *limit*:



The volume and chapter references for *Calculus Illustrated*

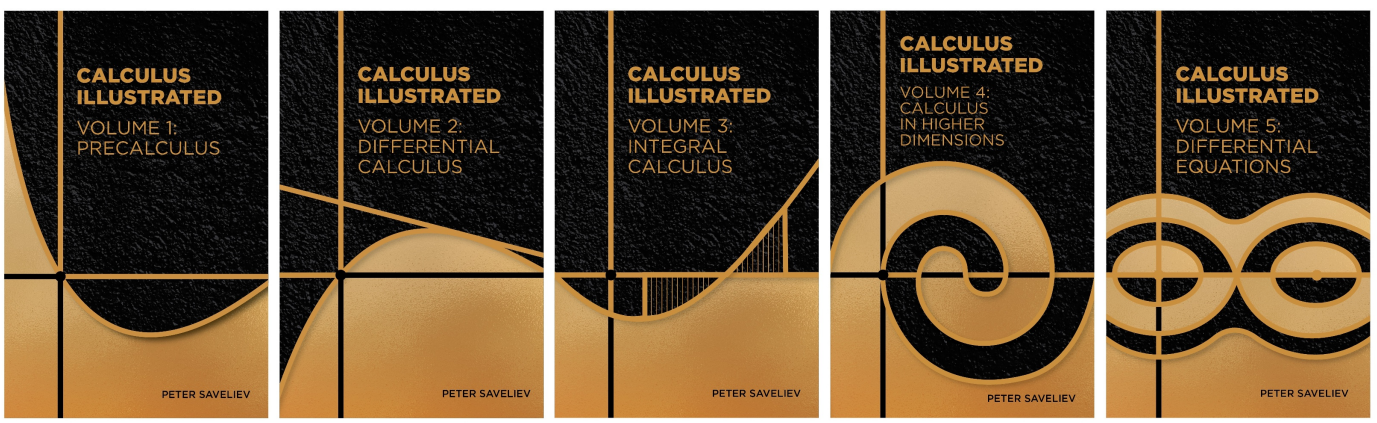
This book is a part of the series *Calculus Illustrated*. The series covers the standard material of the undergraduate calculus with a substantial review of precalculus and a preview of elementary ordinary and partial differential equations. Below is the list of the books of the series, their chapters, and the way the present book (parenthetically) references them.

	■ **Calculus Illustrated. Volume 1: Precalculus**
1 PC-1	Calculus of sequences
1 PC-2	Sets and functions
1 PC-3	Compositions of functions
1 PC-4	Classes of functions
1 PC-5	Algebra and geometry
	■ **Calculus Illustrated. Volume 2: Differential Calculus**
2 DC-1	Limits of sequences
2 DC-2	Limits and continuity
2 DC-3	The derivative
2 DC-4	Differentiation
2 DC-5	The main theorems of differential calculus
2 DC-6	What we can do with calculus
	■ **Calculus Illustrated. Volume 3: Integral Calculus**
3 IC-1	The Riemann integral
3 IC-2	Integration
3 IC-3	Applications of integral calculus
3 IC-4	Several variables
3 IC-5	Series
	■ **Calculus Illustrated. Volume 4: Calculus in Higher Dimensions**
4 HD-1	Functions in multidimensional spaces
4 HD-2	Parametric curves
4 HD-3	Functions of several variables
4 HD-4	The gradient
4 HD-5	The integral
4 HD6	Vector fields
	■ **Calculus Illustrated. Volume 5: Differential Equations**
5 DE-1	Ordinary differential equations
5 DE-2	Vector and complex variables
5 DE-3	Systems of ODEs
5 DE-4	Applications of ODEs
5 DE-5	Partial differential equations

About the author



Peter Saveliev is a professor of mathematics at Marshall University, Huntington, West Virginia, USA. After a Ph.D. from the University of Illinois at Urbana-Champaign, he devoted the next 20 years to teaching mathematics. Peter is the author of a graduate textbook *Topology Illustrated* published in 2016. He has also been involved in research in algebraic topology and several other fields. His non-academic projects have been: digital image analysis, automated fingerprint identification, and image matching for missile navigation/guidance.



Contents

Preface . . . 1

Chapter 1: Calculus of sequences . . . 10
1.1 What is calculus about? . . . 10
1.2 The real number line . . . 20
1.3 Sequences . . . 24
1.4 Repeated addition and repeated multiplication . . . 36
1.5 How to find nth-term formulas for sequences . . . 42
1.6 The algebra of exponents . . . 50
1.7 The Binomial Formula . . . 56
1.8 The sequence of differences: velocity . . . 63
1.9 The sequence of the sums: displacement . . . 72
1.10 Sums of differences and differences of sums: motion . . . 85
1.11 The algebra of sums and differences . . . 91

Chapter 2: Sets and functions . . . 98
2.1 Sets and relations . . . 98
2.2 Functions . . . 106
2.3 Sequences are numerical functions . . . 112
2.4 How numerical functions emerge: optimization . . . 117
2.5 Set building . . . 124
2.6 The xy-plane: where graphs live... . . . 132
2.7 Linear relations . . . 138
2.8 Relations vs. functions . . . 142
2.9 A function as a black box . . . 148
2.10 Give the function a domain... . . . 158
2.11 The graph of a function . . . 165
2.12 Linear functions . . . 171
2.13 Algebra creates functions . . . 179
2.14 The image: the range of values of a function . . . 192

Chapter 3: Compositions of functions . . . 199
3.1 Operations on sets . . . 199
3.2 Piecewise-defined functions . . . 207
3.3 Numerical functions are transformations of the line . . . 217
3.4 Functions with regularities: one-to-one and onto . . . 224
3.5 Compositions of functions . . . 234
3.6 The inverse of a function . . . 244
3.7 Units conversions and changes of variables . . . 259
3.8 Transforming the axes transforms the plane . . . 264
3.9 Changing a variable transforms the graph of a function . . . 273
3.10 The graph of a quadratic polynomial is a parabola . . . 284

Chapter 4: The main classes of functions . . . 291
4.1 The simplest functions . . . 291
4.2 Monotonicity and the extreme values of functions . . . 301
4.3 Functions with symmetries . . . 310

4.4 Quadratic polynomials . . . 321
4.5 The polynomial functions . . . 328
4.6 The rational functions . . . 341
4.7 The root functions . . . 353
4.8 The exponential functions . . . 360
4.9 The logarithmic functions . . . 373
4.10 The trigonometric functions . . . 378

■ Chapter 5: Algebra and geometry . . . 394
5.1 The arithmetic operations on functions . . . 394
5.2 The algebra of compositions . . . 402
5.3 Solving equations . . . 409
5.4 The algebra of logarithms . . . 418
5.5 The Cartesian system for the Euclidean plane . . . 429
5.6 The Euclidean plane: distances . . . 437
5.7 Trigonometry and the wave function . . . 451
5.8 The Euclidean plane: angles . . . 465
5.9 From geometry to calculus . . . 475
5.10 Solving inequalities . . . 480

Exercises . . . 487
1 Exercises: Algebra . . . 487
2 Exercises: Sequences . . . 490
3 Exercises: Sets and logic . . . 491
4 Exercises: Coordinate system . . . 492
5 Exercises: Linear algebra . . . 493
6 Exercises: Polynomials . . . 495
7 Exercises: Relations and functions . . . 497
8 Exercises: Graphs . . . 500
9 Exercises: Compositions . . . 502
10 Exercises: Transformations . . . 505
11 Exercises: Basic models . . . 507

Index . . . 510

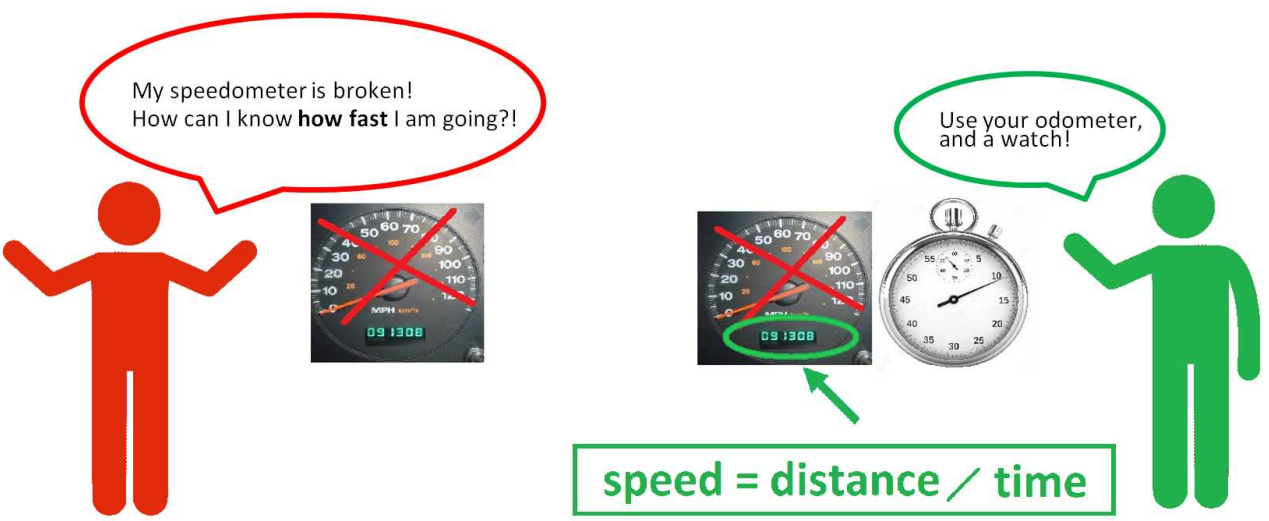
Chapter 1: Calculus of sequences

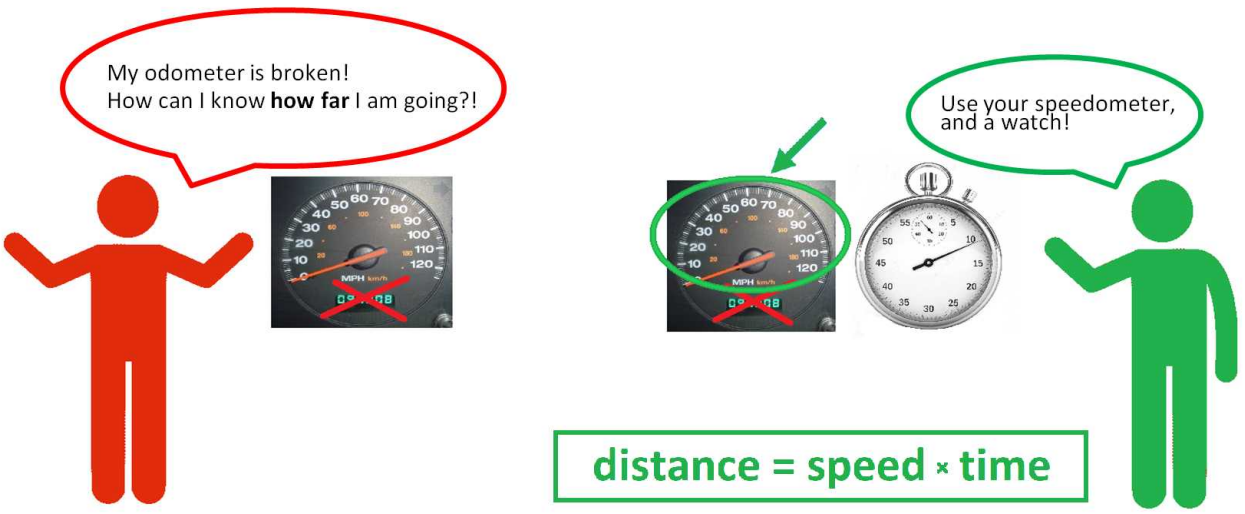
Contents

1.1 What is calculus about? 10
1.2 The real number line 20
1.3 Sequences 24
1.4 Repeated addition and repeated multiplication 36
1.5 How to find nth-term formulas for sequences 42
1.6 The algebra of exponents 50
1.7 The Binomial Formula 56
1.8 The sequence of differences: velocity 63
1.9 The sequence of the sums: displacement 72
1.10 Sums of differences and differences of sums: motion 85
1.11 The algebra of sums and differences 91

1.1. What is calculus about?

We present the idea of calculus in these two related pictures:





The two problems are solved, respectively, with the help of these two versions of the same elementary school formula:

$$\text{speed} = \text{distance} / \text{time} \quad \text{and} \quad \text{distance} = \text{speed} \times \text{time}$$

The equation is solved for the distance or for the speed depending on what is known and what is unknown.

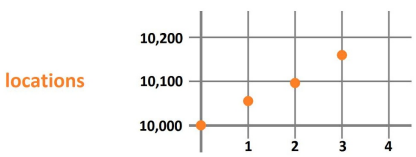
What takes this idea beyond elementary school is the possibility that *velocity varies*.

Let's be more specific. We will face the two situations above but with more data collected and more information derived from it.

First, imagine that our speedometer is broken. What do we do if we want to estimate how fast we are driving? We look at the odometer *several* times – say, every hour on the hour – during the trip and record the mileage on a piece of paper. The list of our consecutive *locations* might look like this:

- initial reading: $10,000$ miles
- after the first hour: $10,055$ miles
- after the second hour: $10,095$ miles
- after the third hour: $10,155$ miles
- etc.

We can plot – as an illustration – the locations against time:



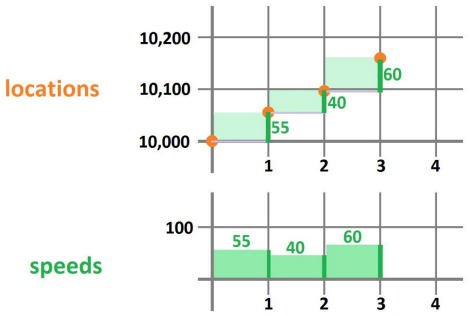
But what do we know about what the speed has been? Nothing without algebra! Fortunately, the algebra is simple:

$$\text{speed} = \frac{\text{distance}}{\text{time}}$$

The time interval was chosen to be 1 hour, so all we need is to find the distance covered during each of these one-hour periods, by *subtraction*:

- distance covered during the first hour: $10,055 - 10,000 = 55$ miles
- distance covered during the second hour: $10,095 - 10,055 = 40$ miles
- distance covered during the third hour: $10,155 - 10,095 = 60$ miles
- etc.

We see below how these new numbers appear as the heights of the steps of our last plot (top):



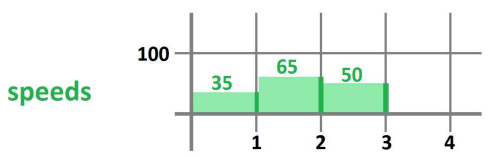
We then plot these new numbers against time (bottom). As you can see, we illustrate the new data in such a way as to suggest that the speed remains *constant* during each of these hour-long periods.

The problem is solved! We have established that the speed has been – roughly – 55, 40, and 60 miles an hour during those three time intervals, respectively.

Now on the flip side: Imagine this time that it is the odometer that is broken. If we want to estimate how far we will have gone, we should look at the speedometer *several* times – say, every hour – during the trip and record its readings on a piece of paper. The result may look like this:

- during the first hour: 35 miles an hour
- during the second hour: 65 miles an hour
- during the third hour: 50 miles an hour
- etc.

Let's plot our speed against time to visualize what has happened:



Once again, we illustrate the data in such a way as to suggest that the speed remains *constant* during each of these hour-long periods.

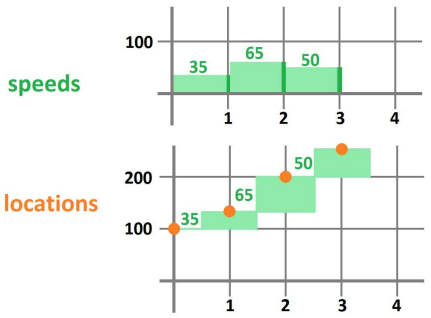
Now, what does this tell us about our location? Nothing, without algebra! Fortunately, we can just use the same formula as before:

$$\text{distance} = \text{speed} \times \text{time}$$

In contrast to the former problem, we need another bit of information. We must know the *starting point* of our trip, say, the 100-mile mark. The time interval was chosen to be 1 hour so that we need only to *add*, and keep adding, the speed at which – we assume – we drove during each of these one-hour periods:

- the location after the first hour: $100 + 35 = 135$-mile mark
- the location after the two hours: $135 + 65 = 200$-mile mark
- the location after the three hours: $200 + 50 = 250$-mile mark
- etc.

In order to illustrate this algebra, we use the speeds as the heights of the consecutive steps of the staircase:



Then the new numbers show how high we have to climb in our last plot.

The problem is solved! We have established that we have progressed through the roughly 135-, 200-, and 250-mile marks during this time.

Our ability to use negative numbers allows us to treat the data the exact same way even when we change directions. As the words "speed" and "distance" imply that these quantities are positive, we speak of "velocity" and "displacement" instead as:

$$\text{velocity} = \frac{\text{displacement}}{\text{time}} \quad \text{and} \quad \text{displacement} = \text{velocity} \cdot \text{time}$$

Changing gears, one of the easier conclusions we derive from these formulas is the following simple statement:

▶ With a positive velocity, we are moving forward.

Indeed, if the time increment is positive and the velocity is positive too, then so is their product, the displacement, according to the second formula.

Let's explore the logic of this statement. It can be recast as an *implication*, i.e., an "if-then" statement:

▶ `IF` the velocity is positive, `THEN` the motion is in the positive direction.

Exercise 1.1.1

Restate as an implication each of the following statements: (a) every square is a rectangle; (b) parallel lines don't intersect; (c) $(a+b)^2 = a^2 + 2ab + b^2$.

We will use the following convenient abbreviation throughout the text:

Implication

$\Longrightarrow$

It reads "then", "therefore", or "implies that".

Then, the above statement takes the following abbreviated form:

▶ The velocity is positive $\Longrightarrow$ the motion is in the positive direction.

Now, we can try to "flip" the implication of this statement, without assuming that the result will be true:

▶ The velocity is positive $\Longleftarrow$ the motion is in the positive direction.

In other words, we have the following implication:

▶ The motion is in the positive direction $\Longrightarrow$ the velocity is positive.

The latter is called the *converse* of the original statement. It's also an implication stated as:

▶ `IF` the motion is in the positive direction, `THEN` the velocity is positive.

The converse is true as well!

This is the abbreviation we used:

Implication

$\Longleftarrow$

It reads "whenever", "provided", or "only if".

Exercise 1.1.2

State the converse of each of the following statements: (a) every square is a rectangle; (b) parallel lines don't intersect; (c) $2x = 2y$ when $x = y$.

Warning!

The converse of a true statement does not have to be true; example: $x = 1 \implies x^2 = 1$.

Exercise 1.1.3

Suggest your own example of a true statement the converse of which is false.

In our case, the implications go both ways! Combined, the statement and its converse form an *equivalence*:

- The velocity is positive `IF AND ONLY IF` the motion is in the positive direction.

We will use the following convenient abbreviation:

Equivalence

$\Longleftrightarrow$

It reads "if and only if" or "is equivalent to".

Then our statement is written as follows:

- The velocity is positive $\Longleftrightarrow$ the motion is in the positive direction.

The two parts of an equivalence are interchangeable!

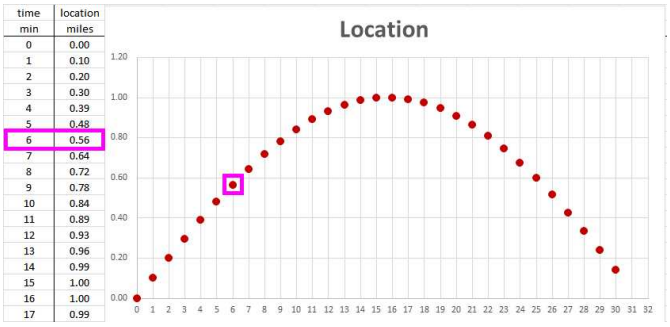
Throughout the text, we will apply this analysis to all statements we make in order to ensure that we know exactly what we are saying and what we are *not* saying.

We next consider more complex examples of the relation between location and velocity. First, *from location to velocity...*

Suppose that this time we have a *sequence* of more than 30 data points (more is indicated by "..."); they are the locations of a moving object recorded every minute:

time	min	0	1	2	3	4	5	6	7	8	9	10	...
location	miles	0.00	0.10	0.20	0.30	0.39	0.48	0.56	0.64	0.72	0.78	0.84	...

This data is also seen in the first two columns of the spreadsheet (left):



time min	location miles
0	0.00
1	0.10
2	0.20
3	0.30
4	0.39
5	0.48
6	0.56
7	0.64
8	0.72
9	0.78
10	0.84
11	0.89
12	0.93
13	0.96
14	0.99
15	1.00
16	1.00
17	0.99

The data is furthermore illustrated as a "scatter plot" (right). It starts to look like a *curve*!

Warning!

The plot is nothing but a visualization of the data.

What has happened to the moving object can now be read from the graph:

- It was moving in the positive direction.
- It was moving fairly fast but then started to slow down.
- It stopped for a very short period.
- It started to move in the opposite direction.
- It started to speed up in that direction.

To find how fast we move over these one-minute intervals, we compute the *differences* of locations for each pair of consecutive locations.

First, the table. We use the data from the row of locations. This is how the first one is computed:

time	min	0	1	...
location	miles	0.00	0.10	...
		↘	↓	
difference			$0.10 - 0.00$	...
			$\|\|$	
velocity	miles/min		0.10	...

We compute this difference for each pair of consecutive locations and then place it in a row for the velocities that we created at the bottom of our table:

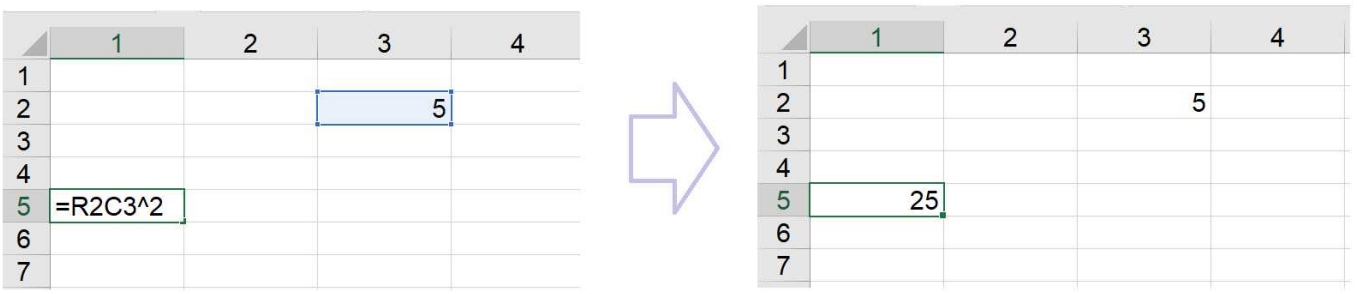
time	min	0	1	2	3	4	5	6	7	8	9	...
location	miles	0.00	0.10	0.20	0.30	0.39	0.48	0.56	0.64	0.72	0.78	...
			↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	↘ ↓	...
velocity	miles/min		0.10	0.10	0.10	0.09	0.09	0.09	0.08	0.07	0.07	...

Example 1.1.4: spreadsheet formulas

We use formulas to pull data from other cells. There are two ways. First, the "absolute" reference:

```
=R2C3^2
```

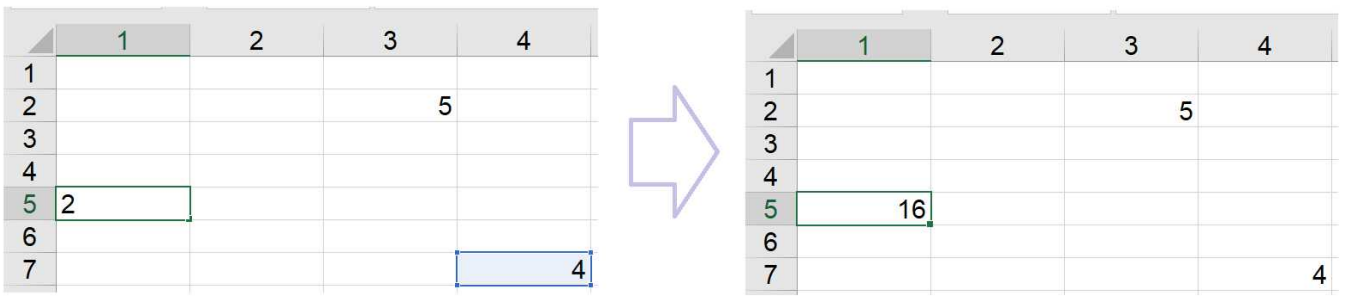
Any cell with this formula will take the value contained in the cell located at row 2 and column 3 and square it:



Second, the “relative" reference:

```
=R[2]C[3]^2
```

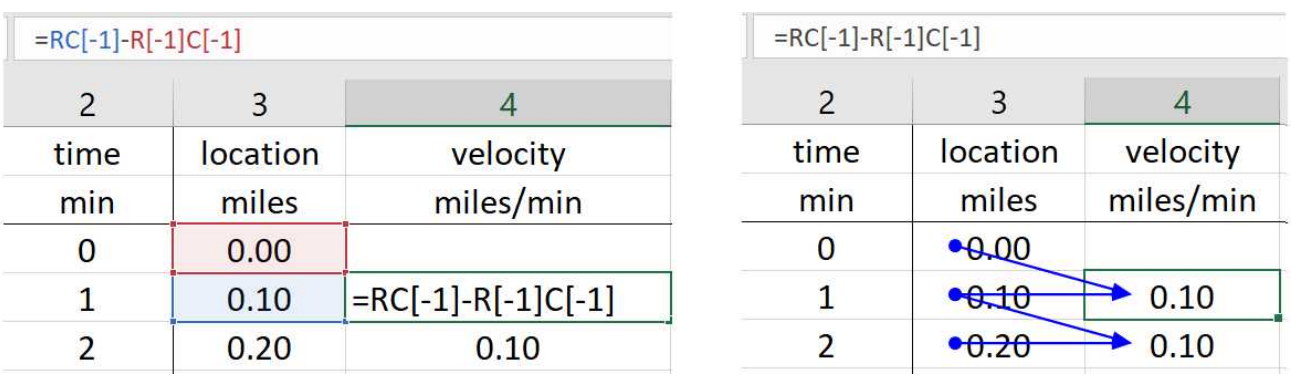
Any cell with this formula will take the value contained in the cell located 2 rows down and 3 columns right from it and square it:



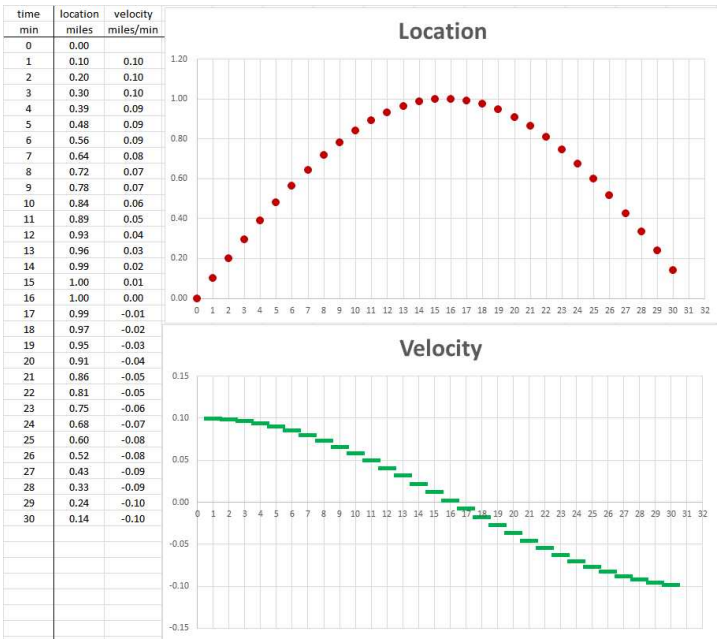
Practically, we'd rather use the spreadsheet. We compute the differences by pulling data from the column of locations with the following formula:

```
=RC[-1]-R[-1]C[-1]
```

Here, the two values come from the last column, `C[-1]`, same row, `R`, and last row, `R[-1]`. Below, you can see the two references in the formulas marked with red and blue (left) and the dependence shown with the arrows (right):



We place the result in a new column we created for the velocities:



time	location	velocity
min	miles	miles/min
0	0.00	
1	0.10	0.10
2	0.20	0.10
3	0.30	0.10
4	0.39	0.09
5	0.48	0.09
6	0.56	0.09
7	0.64	0.08
8	0.72	0.07
9	0.78	0.07
10	0.84	0.06
11	0.89	0.05
12	0.93	0.04
13	0.96	0.03
14	0.99	0.02
15	1.00	0.01
16	1.00	0.00
17	0.99	-0.01
18	0.97	-0.02
19	0.95	-0.03
20	0.91	-0.04
21	0.86	-0.05
22	0.81	-0.05
23	0.75	-0.06
24	0.68	-0.07
25	0.60	-0.08
26	0.52	-0.08
27	0.43	-0.09
28	0.33	-0.09
29	0.24	-0.10
30	0.14	-0.10

This new data is illustrated with the second scatter plot.

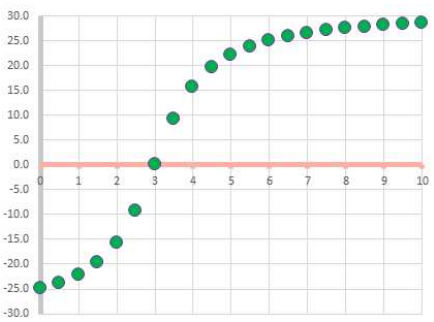
What has happened to the moving object can now be easily read from the second graph:

- The velocity was positive initially (it was moving in the positive direction);
- the velocity was fairly high (it was moving fairly fast) but then it started to decline (slow down);
- the velocity was zero (it stopped) for a very short period;
- then the velocity became negative (it started to move in the opposite direction); and then
- the velocity started to become more negative (it started to speed up in that direction).

Thus, the latter set of data succinctly records some facts about the qualitative and quantitative behavior of the former.

Exercise 1.1.5

The plot of the location is shown below:



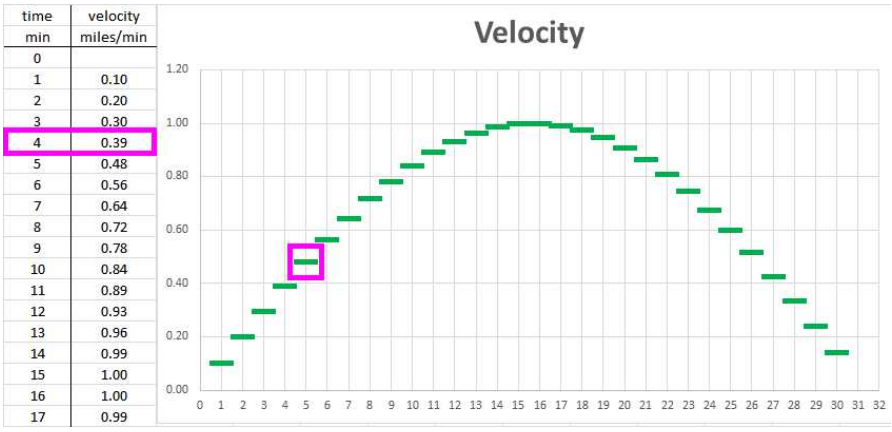
Describe how the velocity and the location have been changing. Sketch the plot of the former.

Now, *from velocity to location...*

Again, we consider 30 data points. These numbers are the values of the velocity of an object recorded every minute:

time	min	0	1	2	3	4	5	6	7	8	9	10	...
velocity	miles/min		0.10	0.20	0.30	0.39	0.48	0.56	0.64	0.72	0.78	0.84	...

This data is also seen in the first two columns of the spreadsheet:



time	velocity
min	miles/min
0	
1	0.10
2	0.20
3	0.30
4	0.39
5	0.48
6	0.56
7	0.64
8	0.72
9	0.78
10	0.84
11	0.89
12	0.93
13	0.96
14	0.99
15	1.00
16	1.00
17	0.99

The data is furthermore illustrated as a scatter plot on the right. Again, we emphasize the fact that the velocity data is referring to time intervals by plotting its values with horizontal bars.

To find out where we are at the end of each of these one-minute intervals, we compute the *sum* of velocity (displacement) for each interval by pulling the data from the row of velocities with the previous result. This is how the first one is computed, under the assumption that the initial location is 0:

time	min	0	1	...
velocity	miles/min		0.10	...
			↓	
sum		0.00+	0.10	...
		↑	||	
location	miles	0.00	0.10	...

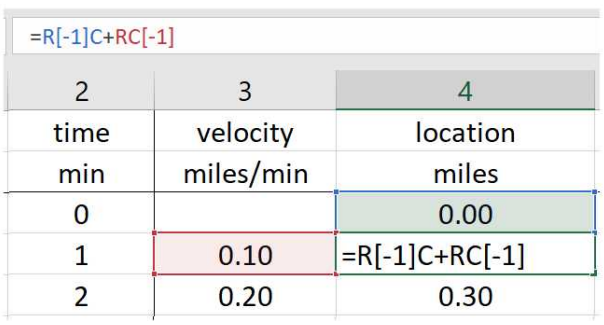
We place this data in a new row added to the bottom of our table:

time	min	0	1	2	3	4	5	6	7	8	...
velocity	miles/min		0.10	0.20	0.30	0.39	0.48	0.56	0.64	0.72	...
			↓	↓	↓	↓	↓	↓	↓	↓	...
location	miles	0.00 →	0.10 →	0.30 →	0.59 →	0.98 →	1.46 →	2.03 →	2.67 →	3.39 →	...

Practically, we use the spreadsheet. We compute the sums by pulling the data from the column of velocities using the following formula:

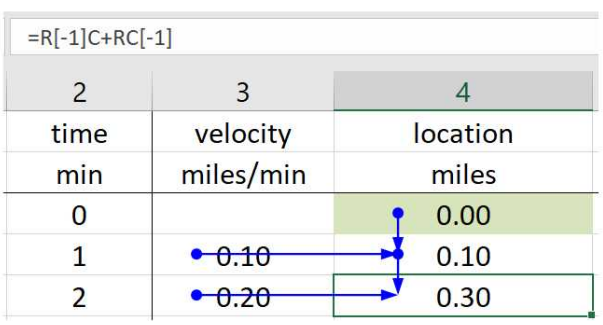
```
=R[-1]C+RC[-1]
```

Here, the two values come from the same, `C`, or last, `C[-1]`, column and the same, `R`, and last, `R[-1]`, row, as follows:



=R[-1]C+RC[-1]

2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	=R[-1]C+RC[-1]
2	0.20	0.30

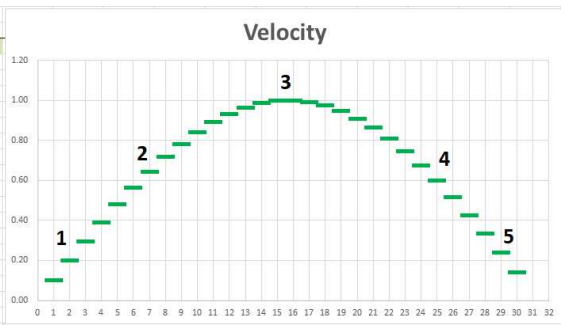


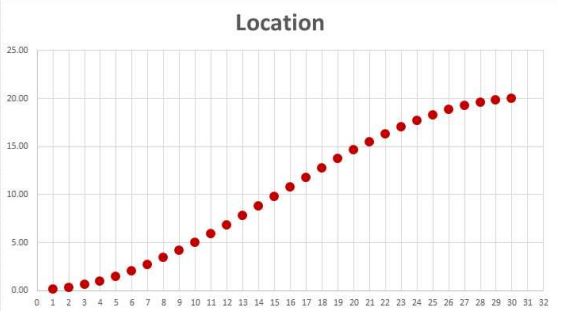
=R[-1]C+RC[-1]

2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	0.10
2	0.20	0.30

We place the result in a new column for locations:

time min	velocity miles/min	location miles
0		0.00
1	0.10	0.10
2	0.20	0.30
3	0.30	0.59
4	0.39	0.98
5	0.48	1.46
6	0.56	2.03
7	0.64	2.67
8	0.72	3.39
9	0.78	4.17
10	0.84	5.01
11	0.89	5.91
12	0.93	6.84
13	0.96	7.80
14	0.99	8.79
15	1.00	9.78
16	1.00	10.78
17	0.99	11.77
18	0.97	12.75
19	0.95	13.70
20	0.91	14.60
21	0.86	15.47
22	0.81	16.28
23	0.75	17.02
24	0.68	17.70
25	0.60	18.30
26	0.52	18.81
27	0.43	19.24
28	0.33	19.57
29	0.24	19.81
30	0.14	19.95



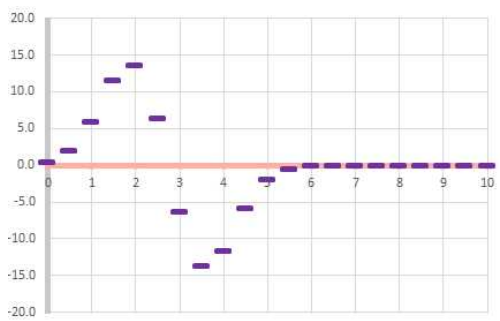


The data is also illustrated as the second scatter plot on the right.

Thus, as the former data set records some facts about the quantitative behavior of the latter, we are able to combine this information to recover the latter.

Exercise 1.1.6

The plot of the velocity is shown below:



Describe how the velocity and the location have been changing. Act it out by stepping left and right. Sketch the plot of the location.

This is what we have discovered:

- *We can tell the velocity from the location and the location from the velocity.*

Let's make sure we know what we are and are not saying. We *are* saying:

- `IF` we know the location, `THEN` we know the velocity.

We are *not* saying the converse. This would be false: We know the velocity $\not\Longrightarrow$ we know the location. The correct, "partial", converse is the following:

- `IF` we know the velocity `AND` we know the initial location, `THEN` we know the location.

Exercise 1.1.7

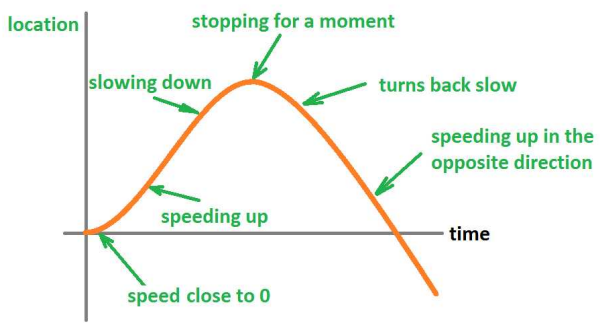
(a) Your location is recorded every half-hour, shown below. Estimate your velocity as it changes with time. (b) Your velocity is recorded every half-hour, shown below. Estimate your location as it changes

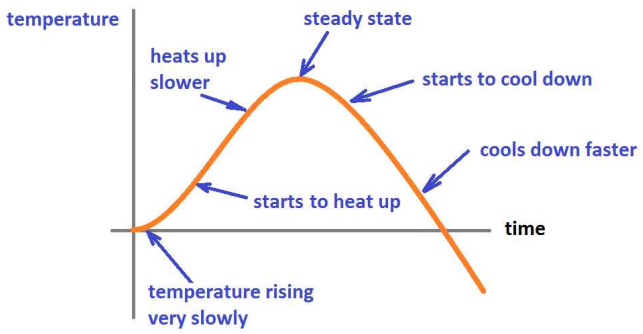
with time.

time	location
0	20
.5	30
1	20
1.5	20
2	50

time	velocity
0	5
.5	3
1	10
1.5	10
2	−10

We can continue to increase the number of data points and, as we zoom out, the scatter plots will look like *continuous curves*! The analysis presented in this section remains fully applicable. It amounts to looking at how fast the vertical location is changing relative to the change of the horizontal location (left):





Furthermore, we replace (right) our time-dependent quantity, the location, for another, the temperature as a single example of the breadth of applicability of these ideas.

1.2. The real number line

The starting point of studying numbers is the *natural numbers*:

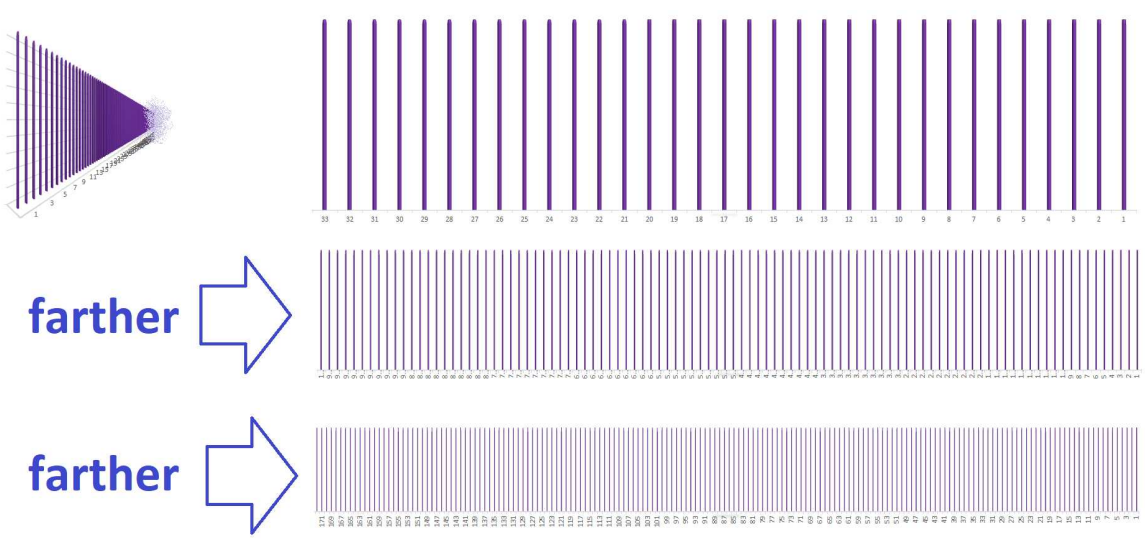
$$0, 1, 2, 3, ...$$

They are initially used for counting. The next step is the *integers*:

$$..., -3, -2, -1, 0, 1, 2, 3, ...$$

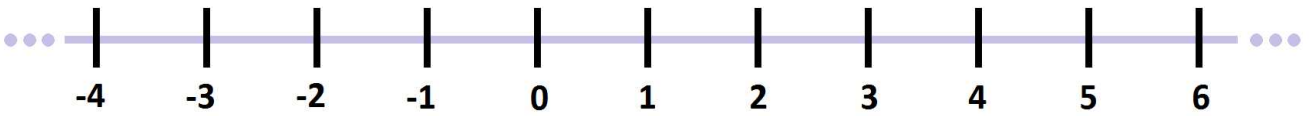
They can be used for studying the space and locations, as follows.

Imagine facing a fence so long that you can't see where it ends. We step *away* from the fence multiple times and there is still more to see:



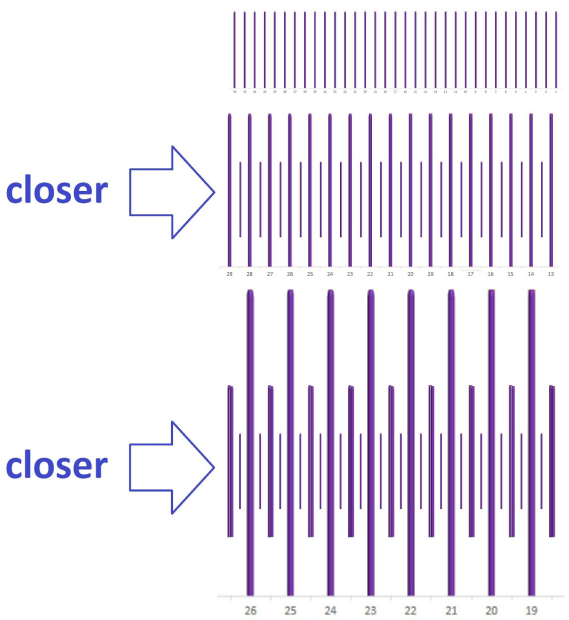
Is the number of planks *infinite*? It may be. For *convenience*, we will just assume that we can go on with this for as long as necessary.

We visualize these as markings on a straight line, according to the order of the planks:



The assumption is that the line and the markings continue without stopping in both directions, which is commonly represented by "...". The same idea applies to the *milestones* on the road; they are also ordered and might continue indefinitely.

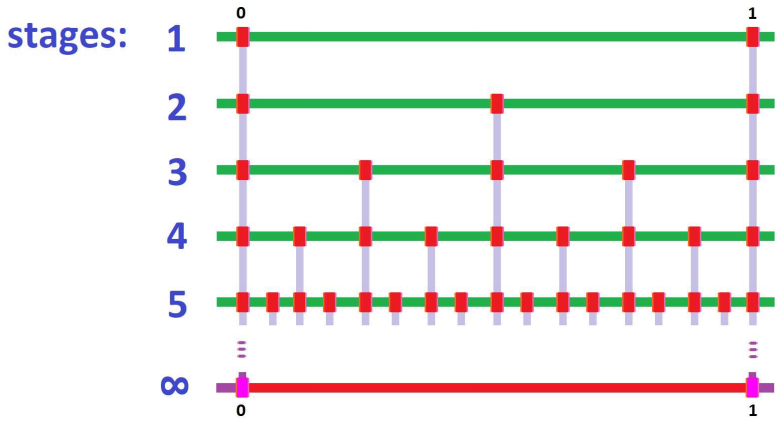
So, we zoomed *out* to see the fence. Suppose now we zoom *in* on a location on the fence. What if there is a shorter plank between every two planks? We look closer and we see more:



If we keep zooming in, the result will look similar to a *ruler*:

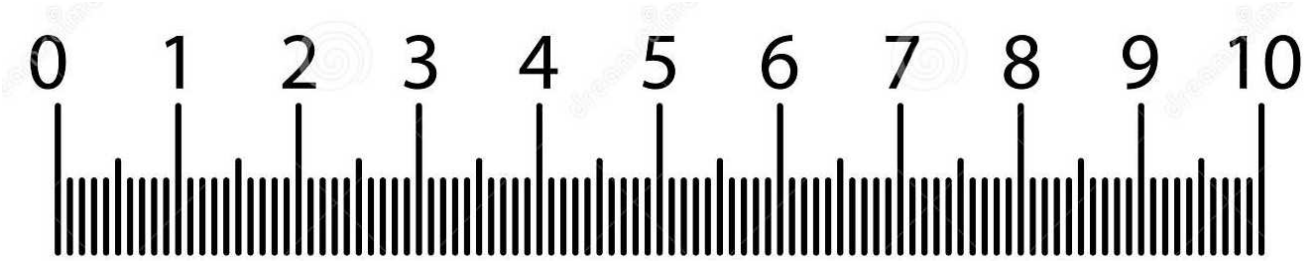


It's as if we add *one mark* between any two and then add another one between either of the two pairs we have created. We keep repeating this step. Even though this ruler goes only to 1/16 of an inch, we can imagine that the process continues indefinitely:



Is the depth *infinite*? It may be. For *convenience*, we will just assume that we can go on with this for as long as necessary.

If we add *nine marks* at a time, the result is a *metric ruler*:



Here, we go from meters to decimeters, to centimeters, to millimeters, etc.

To see it another way, we allow more and more decimals in our numbers:

1.55 :	1.	1.5	1.55	1.550	1.5500	...
1/3 :	.3	.33	.333	.3333	.33333	...
π :	3.	3.1	3.14	3.141	3.1415	...

In order to visualize *all* numbers, we first arrange the integers in a line and then the *line of numbers* is built. It takes several steps.

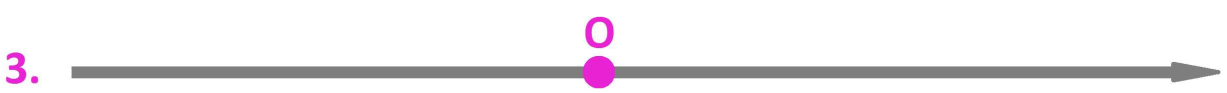
Step 1: Draw a line, called an *axis* (horizontal when convenient):



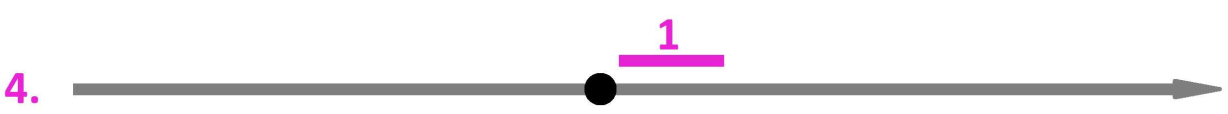
Step 2: Choose one of the two ends of the line as the *positive* direction (the one on the right when convenient), then the other is the *negative*:



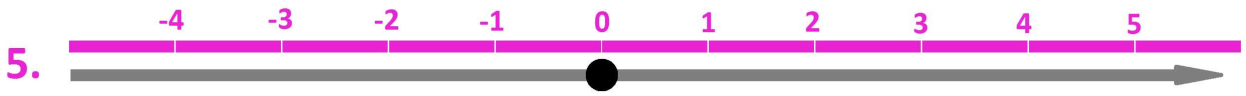
Step 3: Set a point O (a letter, not a number) as the *origin*:



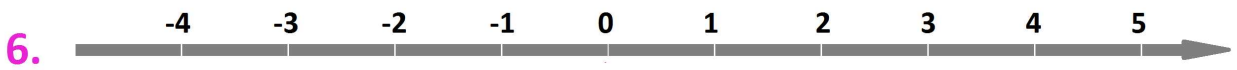
Step 4: Choose a segment of the line as the *unit* of length:



Step 5: Use the segment to measure distances to locations from the origin O (positive in the positive direction, and negative in the negative direction) and add marks, called the *coordinates*:



Step 6: Divide the segments further into fractions of the unit, etc.:

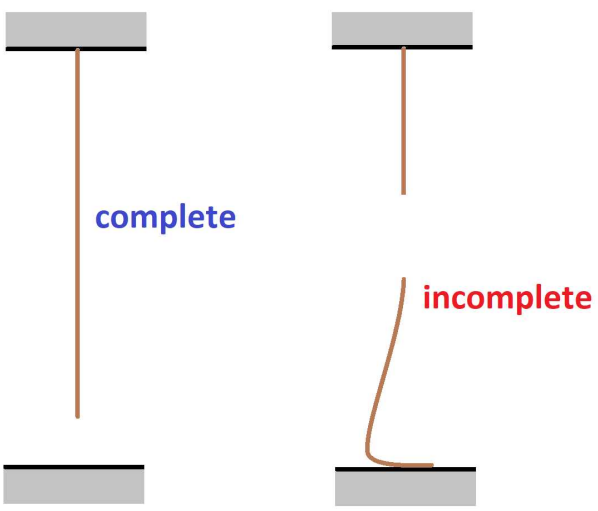


The end result depends on what the building block is. It may contain gaps and look like a ruler (or a comb) as discussed above. It may also be solid and look like a tile or a domino piece:



So, we start with integers as locations and then – by cutting these intervals further and further – also include fractions, i.e., *rational numbers*.

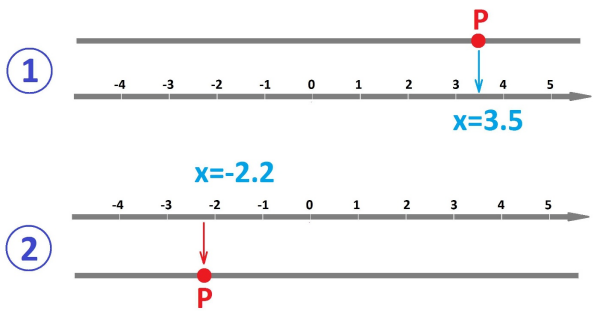
However, we then realize that some of the locations have no counterparts among these numbers. For example, $\sqrt{2}$ is the length of the diagonal of a 1×1 square (and a solution of the equation $x^2 = 2$); it's not rational. That's how the *irrational numbers* come into play. Together, they make up the *real numbers* and the *real number line*. We think of this line as *complete*; there are no missing points. As an illustration, an "incomplete" rope won't hang:



We use this setup to produce a correspondence between the locations on the line and the real numbers:

location P $\longleftrightarrow$ number x

We will follow this correspondence in *both directions*, as follows:



1. First, suppose P is a *location* on the line. We then find the corresponding mark on the line. That's the "coordinate" of P: some *number* x.
2. Conversely, suppose x is a *number*. We think of it as a "coordinate" and find its mark on the line. That's the *location* of x: some point P on the line.

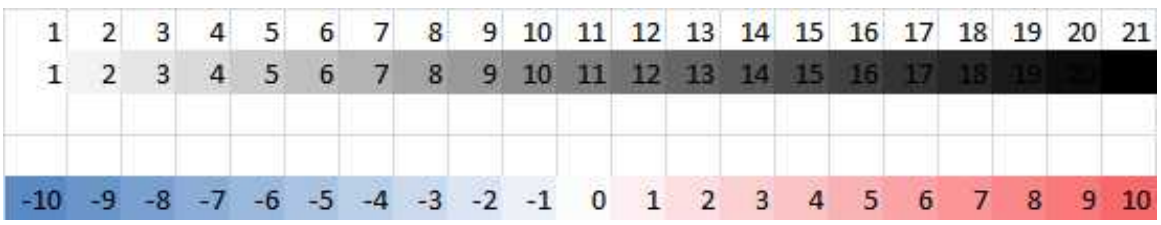
Once this system of coordinates is in place, it is acceptable to think of every location as a number, and vice versa. In fact, we often write:

$$P = x\,.$$

The result may be described as the "1-dimensional coordinate system". It is also called the *real number line* or simply *the number line*.

We have created a visual model of the real numbers. Depending on the real number or a collection of numbers that we are trying to visualize, we choose what part of the real line to exhibit; for example, the zero may or may not be in the picture. We also have to choose an appropriate length of the unit segment in order for the numbers to fit in.

In addition to the ruler, another way to visualize numbers is with *colors*. In fact, in digital imaging the levels of gray are associated with the numbers between 0 and 255. A shorter scale – $1, 2, ..., 20$ – is used in the illustration below (top):



It is also often convenient to associate blue with negative and red with positive numbers (bottom).

Exercise 1.2.1

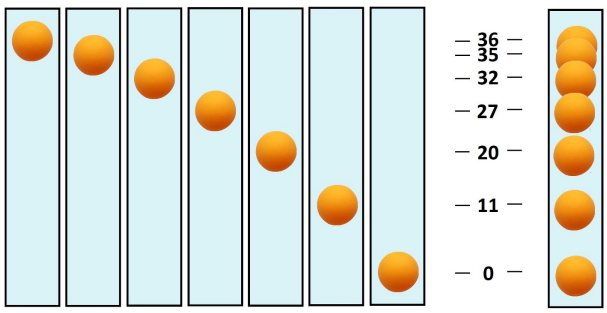
Think of other examples when numbers are visualized with colors.

1.3. Sequences

The lists of numbers considered in the first section are *sequences*: the locations and the velocities.

Example 1.3.1: falling ball

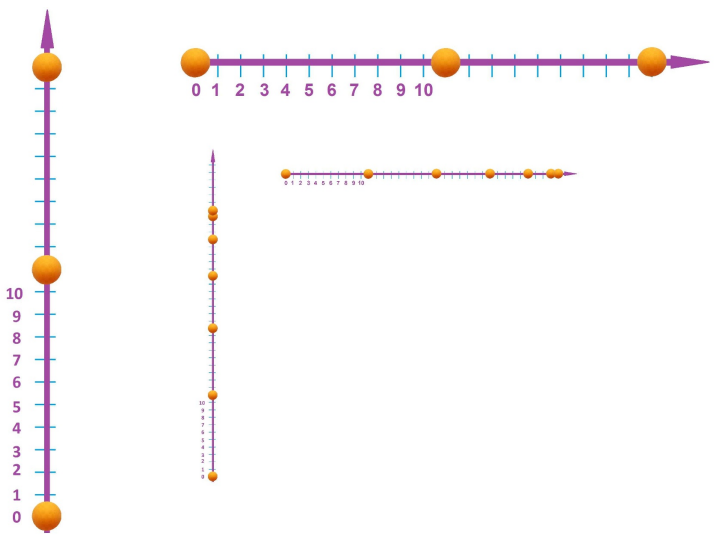
We videotape a ping-pong ball falling down and record – at equal intervals – how high it is. The result is an ever-expanding string, a sequence, of numbers. If the frames of the video are combined into one image, it will look something like this:



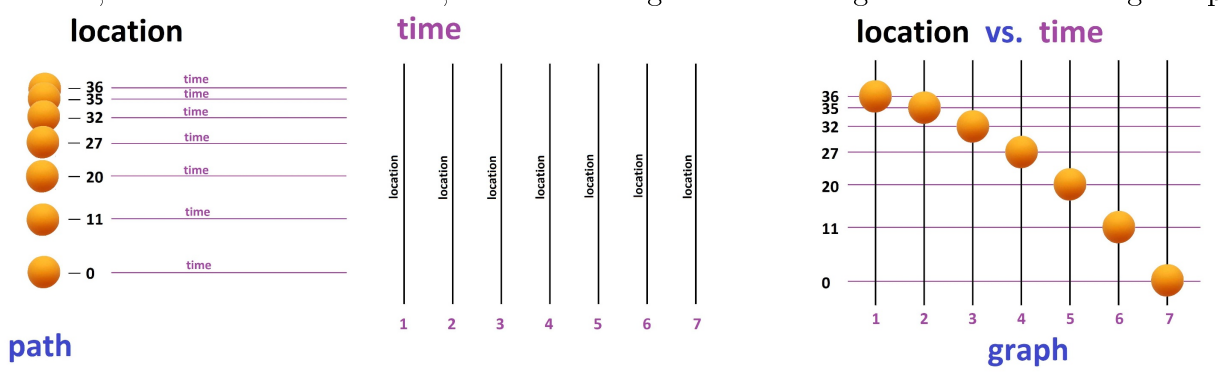
We ignore the time for now and concentrate on the locations only. We have the first few in a *list*:

$$36,\ 35,\ 32,\ 27,\ 20,\ 11,\ 0\,.$$

This data can be visualized by placing the ball at every coordinate location on the real line, oriented vertically or horizontally:



Though not uncommon, this method of visualization of motion, or of sequences in general, has its drawbacks: Overlapping may be inevitable and the *order* of events is lost (unless we add labels). A more popular approach is the following. The idea is to *separate time and space*, to give a separate real line, an axis, to each moment of time, and then bring them back together in one rectangular plot:



The location varies – as it does – vertically while the time progresses horizontally. The result is similar to the collection of the frames of the video as seen above. The plot is called the *graph* of the sequence. As far as the data is concerned, we have a list of *pairs*, time and location, arranged in a table:

moment	height
1	36
2	35
3	32
4	27
5	20
6	11
7	0

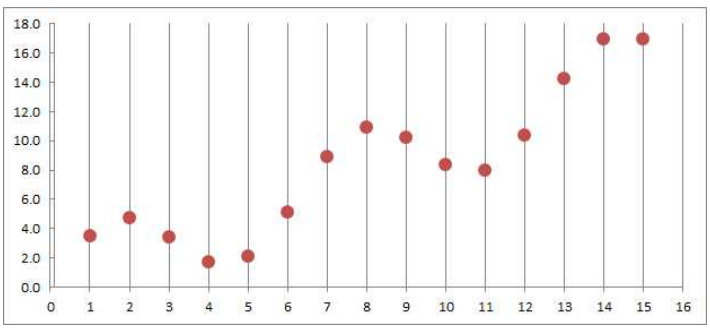
The table is just as effective representation of the data if we flip it; it's more compact:

moment:	1	2	3	4	5	6	7
height:	36	35	32	27	20	11	0

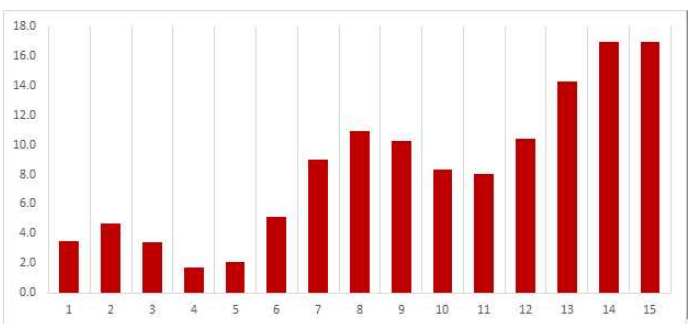
Warning!

It is entirely a matter of convenience to represent our data as a two-column table (especially in a spreadsheet) or a two-row table. In either case, it's a list of pairs of numbers.

So, the most common way to visualize a sequence of *numbers* is as a sequence of *points* on a sequence of vertical axes:



It is also common to represent the same numbers as vertical *bars*:



Warning!

The graph is just a visualization of the data.

To represent a sequence algebraically, we first give it a name, say, a, and then assign a specific variation of this name to each term of the sequence:

Indices of sequence

index: n	1	2	3	4	5	6	7	...
term: a_n	a_1	a_2	a_3	a_4	a_5	a_6	a_7	...

The *name* of a sequence is a letter, while the subscript called the *index* indicates the place of the term within the sequence. It reads: "a sub 1", "a sub 2", etc.

The letter "n" is often the preferred choice for the index because it might stand for "natural numbers": $1, 2, 3, 4, \ldots$. As before, "..." indicates a continuing pattern: The indices continue to grow incrementally.

Example 1.3.2: falling ball

For the last example, let's name the sequence h for "height". Then the above *table* take this form:

moment:	1	2	3	4	5	6	7	...
height:	h_1	h_2	h_3	h_4	h_5	h_6	h_7	...
	||	||	||	||	||	||	||	...
height:	36	35	32	27	20	11	0	...

This is the same table aligned vertically:

moment	height	height
1	$h_1 =$	36
2	$h_2 =$	35
3	$h_3 =$	32
4	$h_4 =$	27
5	$h_5 =$	20
6	$h_6 =$	11
7	$h_7 =$	0
..	..	..

Either table is a list of identities that can be written in any order:

$$h_3 = 32 \quad h_5 = 20 \quad h_7 = 0 \quad h_4 = 27$$

$$h_1 = 36 \quad h_2 = 35 \quad h_4 = 27 \quad h_6 = 11$$

Let's deconstruct the notation:

Index of a term

$$\underset{\text{name}}{\uparrow}\ a_{\ \downarrow\ \underset{\text{index}}{n}}$$

In other words, we specify a sequence first and then specify the location of the term within the sequence.

Indices serve as *tags*:



A sequence can come from a list or a table unless it's infinite. Infinite sequences often come from *formulas*.

Example 1.3.3: sequence of reciprocals

The formula:

$$a_n = 1/n\,,$$

gives rise to the sequence,

$$a_1 = 1,\ a_2 = 1/2,\ a_3 = 1/3,\ a_4 = 1/4,\ \ldots$$

Indeed, replacing n in the formula with 1, then 2, 3, etc. produces the numbers on the list one by one, as follows. We enter n into the formula, and a_n appears at the end of the computation. In other words, we place the current value of n inside a blank box (where n used to be) in the formula:

$$\underset{\text{insert } n}{a_{\square}} \quad = \quad \frac{1}{\underset{\text{insert } n}{\square}}.$$

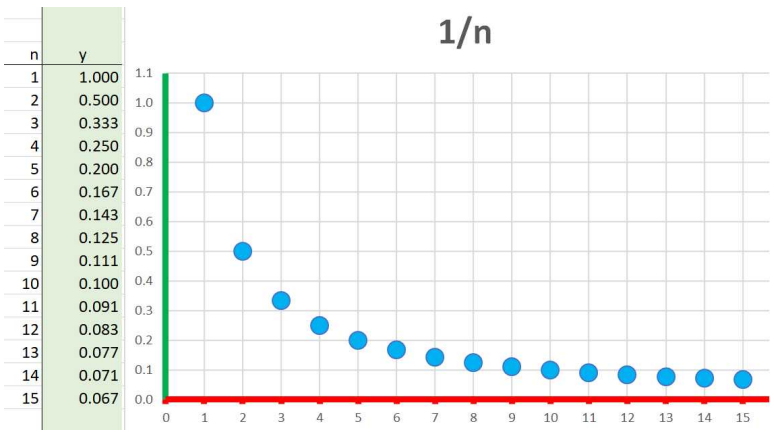
It is called "substitution". We do this seven times below:

n	1	2	3	4	5	6	7	...
a_n	a_1	a_2	a_3	a_4	a_5	a_6	a_7	...
$\|\|$	$\|\|$	$\|\|$	$\|\|$	$\|\|$	$\|\|$	$\|\|$	$\|\|$	...
$\frac{1}{n}$	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	...

With a formula, we can use a spreadsheet (a vertical table) to produce more values with the formula:

```
=1/RC[-1]
```

We also plot these values:



The complete, algebraic, representation is as follows:

$$a_n = 1/n, \ n = 1, 2, 3, ...$$

Exercise 1.3.4

Write a few terms of the sequences given by the formulas:

1. $a_n = 3n - 1$
2. $b_n = 1 + \dfrac{1}{n}$

We will say that this is the *nth-term formula of the sequence.*

Below is the simplest kind.

Definition 1.3.5: constant sequence

A *constant sequence* has all its terms equal to each other.

In other words, we have

$$a_1 = a_2 = ... = c$$

for some number c.

Thus, *every* formula is capable of creating an infinite sequence a_n. For example, we can take these:

- $a_n = n$
- $b_n = n^2$
- $c_n = n^3$
- etc.

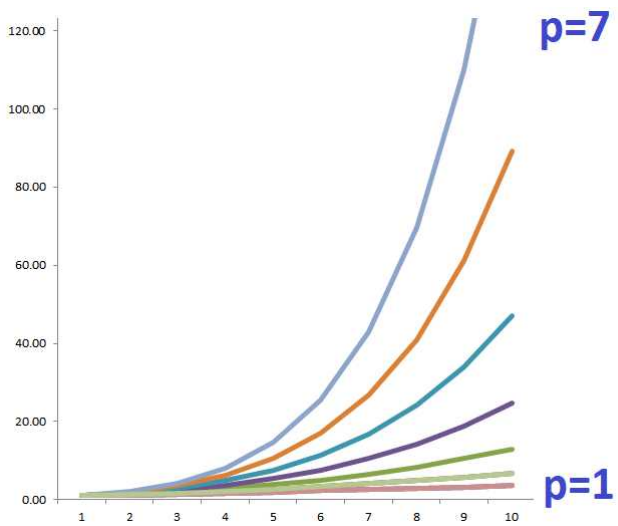
They make up a whole class of sequences.

Definition 1.3.6: power sequence

For every positive integer p, a *power sequence*, or a p-sequence, is given by the formula:

$$a_n = n^p, \quad n = 1, 2, 3, ...$$

The relation between the sequences becomes clear if we zoom out from their graphs ($p = 1, 2, ..., 7$):

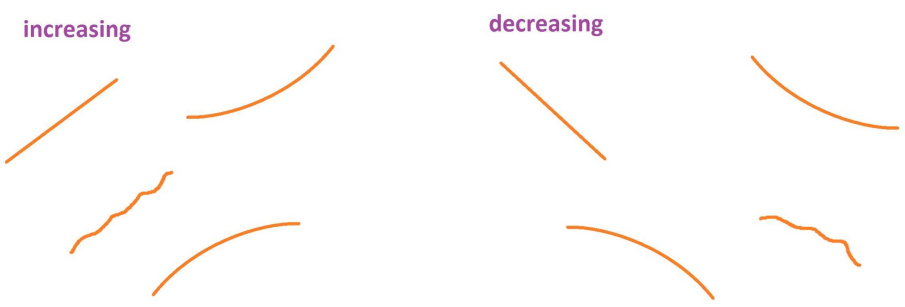


Indeed, the larger the power p, the faster the sequence grows.

The relation between the consecutive terms of each sequence is also clear: It grows! We use these words:

- "growth" or "increase" when we see the graph that *rises* left to right, and
- "decline" or "decrease" when we see the graph that *drops* left to right,

as follows:



As you can see, the behavior varies even within these two categories.

The precise definition has to rely on considering *every pair of consecutive terms* of the sequence. For example, the sequence of the falling ball,

$$36,\ 35,\ 32,\ 27,\ 20,\ 11,\ 0\,,$$

is decreasing because

$$36 > 35 > 32 > 27 > 20 > 11 > 0\,.$$

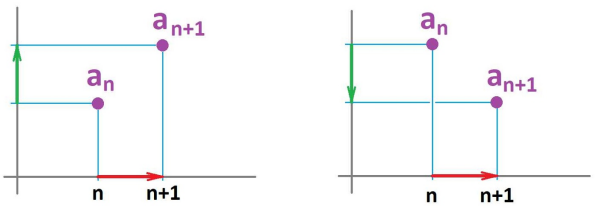
For the general case, we write:

$$a_1 > a_2 > a_3 > a_4 > a_5 > a_6\,.$$

In other words, the current term, a_n, is larger than the next, a_{n+1}:

$$\boxed{\text{last} \;\geq\; \text{next}}$$

The meaning of this inequality is clear when we zoom in on the graph (right):



On left, the sequence is increasing:

$$\boxed{\text{next} \;\geq\; \text{last}}$$

The following will be used throughout.

Definition 1.3.7: increasing sequence and decreasing sequence

- A sequence a_n is called *increasing* if, for all n, we have

$$a_n \leq a_{n+1}\,.$$

- A sequence is called *decreasing* if, for all n, we have

$$a_n \geq a_{n+1}\,.$$

Collectively, they are called *monotone*.

Warning!

Both increasing and decreasing sequences may have segments with no change; furthermore, a constant sequence is both increasing and decreasing.

Example 1.3.8: proving monotonicity

When the sequence is given by its formula, we use it directly. The sequence $a_n = n^2$ is proven to be increasing as follows. We need to show that for all n we have:

$$n^2 < (n+1)^2\,.$$

We simply expand the right-hand side:

$$(n+1)^2 = n^2 + 2n + 1\,.$$

As n is positive, the last part, $2n+1$, is positive too. Therefore, the expression is larger than n^2.

Similarly, we show that $\dfrac{1}{n}$ is decreasing via the following algebra:

$$n < n+1 \implies \frac{1}{n} > \frac{1}{n+1}\,.$$

The sequence

$$1, -1, 1, -1, 1, -1, 1, -1, 1, -1, ...$$

is neither increasing nor decreasing, i.e., it's not monotone.

Exercise 1.3.9

Show that $\dfrac{1}{n^2}$ is decreasing.

Exercise 1.3.10

Show that all power sequences are increasing.

A major reason why we study sequences is that, in addition to tables and formulas, a sequence can be defined by computing its terms in a *consecutive manner*, one at a time.

Example 1.3.11: regular deposits

A person starts to deposit \$20 every month in his bank account that already contains \$1000. Then, after the first month the account contains:

$$\$1000 + \$20 = \$1020\,,$$

after the second:

$$\$1020 + \$20 = \$1040\,,$$

and so on. In other words, we have:

$$\text{next} \;=\; \text{last} \;+ 20\,.$$

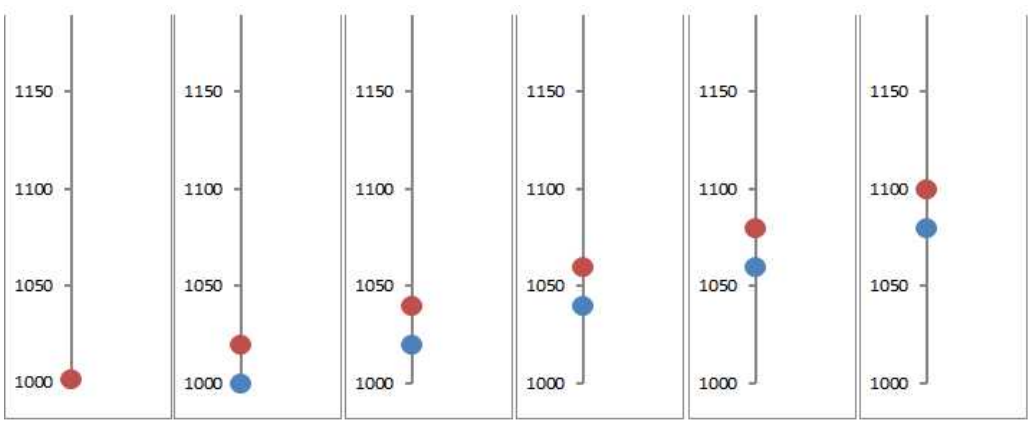
Let's make this algebraic. Suppose a_n is the amount in the bank account after n months, then we have a formula for this sequence:

$$a_{n+1} = a_n + 20\,.$$

How much will he have after 50 years? We'd have to carry out $50 \cdot 12 = 600$ additions. For the spreadsheet, the formula refers to the last row and adds 20, as follows:

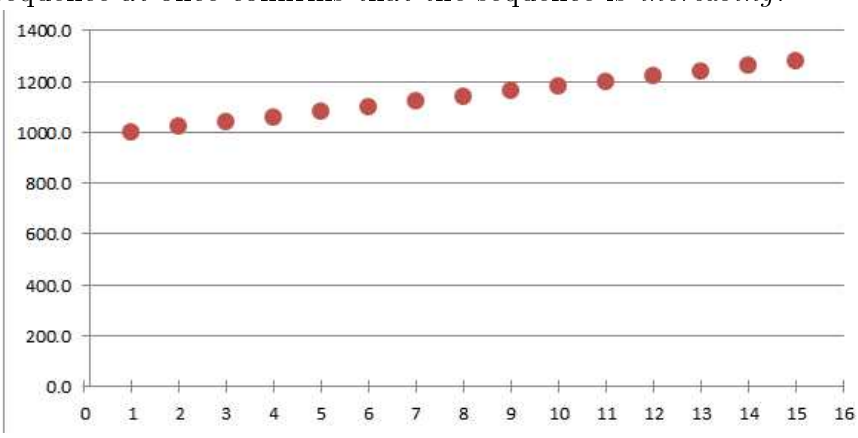
```
=R[-1]C+20
```

Below, the current amount is shown in blue and the next – computed from the current – is shown in red:



Plotting several terms of the sequence at once confirms that the sequence is *increasing*:

time	amount
1	1000.00
2	1020.00
3	1040.00
4	1060.00
5	1080.00
6	1100.00
7	1120.00
8	1140.00
9	1160.00
10	1180.00
11	1200.00
12	1220.00
13	1240.00
14	1260.00
15	1280.00



It also looks like a straight line.

Definition 1.3.12: recursive sequence

We say that a sequence is *recursive* when its next term is found from the current term by a specified formula, i.e., a_n determines a_{n+1}.

This is the difference between computing a sequence directly, such as $a_n = n^2$, and recursively, such as $a_{n+1} = a_n + 20$:

n		a_n
1	$\to$	a_1
2	$\to$	a_2
3	$\to$	a_3
..		..

n	a_n
1	a_1
	$\downarrow$
2	a_2
	$\downarrow$
3	a_3
	$\downarrow$
..	..

The following will be routinely used.

Definition 1.3.13: arithmetic progression

A sequence defined (recursively) by the formula:

$$a_{n+1} = a_n + b$$

is called an *arithmetic progression* with b as its *increment*.

Exercise 1.3.14

If the increment is zero, the sequence is...

Example 1.3.15: compounded interest

An arithmetic progression describes a repetitive process. Also repetitive is the following typical situation. A person deposits \$1000 in his bank account that pays 1% APR compounded annually. Then, after the first year, the interest is

$$\$1000 \cdot .01 = \$10\,,$$

and the total amount becomes \$1010. After the second year, the interest is

$$\$1010 \cdot .01 = \$10.10\,,$$

and so on. In other words, the total amount is multiplied by .01 at the end of each year and then added to the total. An even simpler way to algebraically describe this is to say that the total amount is multiplied by 1.01 at the end of each year, as follows. After the first year, the total is equal to

$$\$1000 \cdot 1.01 = \$1010\,.$$

After the second year, the total is equal to

$$\$1010 \cdot 1.01 = \$1020.1\,,$$

and so on. In other words, we have:

$$\text{next} \;=\; \text{last} \;\cdot 1.01\,.$$

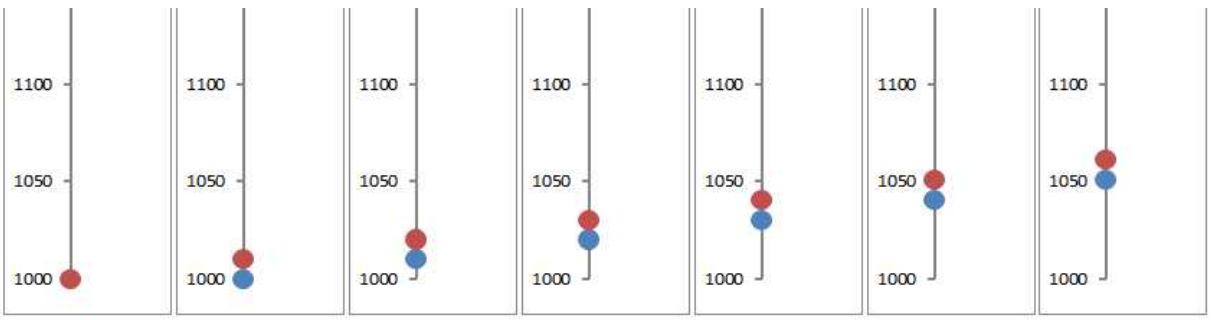
Let's make this algebraic. Suppose a_n is the amount in the bank account after n years. Then we have the following *recursive formula*:

$$a_{n+1} = a_n \cdot 1.01\,.$$

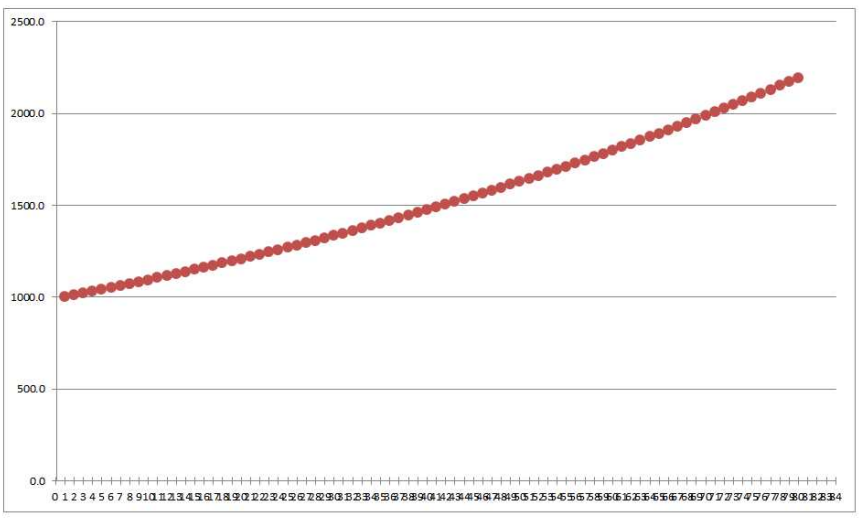
How much will he have after 50 years? We'd have to carry out 50 multiplications. For the spreadsheet, the formula refers to the last row (`R[-1]`) and multiplies by 1.01, as follows:

```
=R[-1]C*1.01
```

We plot a term and the next one:



Only after repeating the step 100 times can one see that this isn't just a straight line:



The sequence is increasing.

The following will be routinely used.

Definition 1.3.16: geometric progression

A sequence defined (recursively) by the formula:

$$a_{n+1} = a_n \cdot r\,,$$

with $r \neq 0$, is called a *geometric progression* with r as its *ratio*. We say that this is:

- a *geometric growth* when $r > 1$, and
- a *geometric decay* when $r < 1$.

Alternatively, it is called an *exponential* growth and decay, respectively.

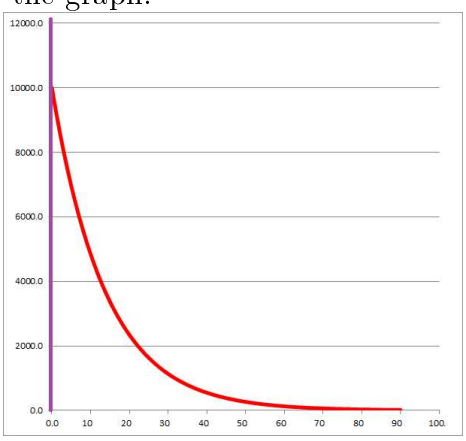
Example 1.3.17: population loss

If the population of a city declines by 3% every year, it is left with 97% of its population at the end

of each year. The result is found by multiplying by .97, every time. We have, therefore:

$$\overbrace{\underbrace{\underbrace{(1,000,000 \cdot 0.97)}_{\text{after 1 year}} \cdot 0.97}_{\text{after 2 years}} \cdot 0.97}^{\text{after 3 years}} .$$

And so on. What will be the population after 50 years? We'd have to carry out 50 multiplications. The long-term trend is clear from the graph:



This is a geometric progression with ratio $r = .97$, i.e., a geometric decay. The sequence is decreasing and eventually there is almost nobody left.

Example 1.3.18: deposits and interest, together

What if we deposit money to our bank account *and* receive interest? The recursive formula is simple, for example:

$$a_{n+1} = a_n \cdot 1.05 + 2000 .$$

Here, the interest is 5% with a \$2000 annual deposit.

Exercise 1.3.19

What does a 2% inflation do to a dollar hidden in the mattress?

Any algebraic operation, or several operations together, can produce a recursive sequence:

$$a_0 \to \boxed{\text{add 2}} \to a_1 \to \boxed{\text{add 2}} \to \ldots \to \boxed{\text{add 2}} \to a_n \to \ldots$$

$$a_0 \to \boxed{\text{divide by 3}} \to a_1 \to \boxed{\text{divide by 3}} \to \ldots \to \boxed{\text{divide by 3}} \to a_n \to \ldots$$

$$a_0 \to \boxed{\text{square it}} \to a_1 \to \boxed{\text{square it}} \to \ldots \to \boxed{\text{square it}} \to a_n \to \ldots$$

Exercise 1.3.20

How do these recursive sequences depend on the value of a_0?

When a sequence is defined recursively, we'd need to carry out this definition n times in order to find the nth term. This is in contrast to sequences defined directly via its *nth-term formula*, such as $a_n = n^2$, that requires a single computation to find any term.

Below, we keep multiplying, just as in the geometric progression, but this time the multiple grows.

Definition 1.3.21: factorial

The *factorial* is the sequence defined (recursively) as follows:

$$a_0 = 1, \quad a_n = a_{n-1} \cdot n, \ n \geq 1 ;$$

i.e.,

$$a_n = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n\,.$$

It is denoted by the following:

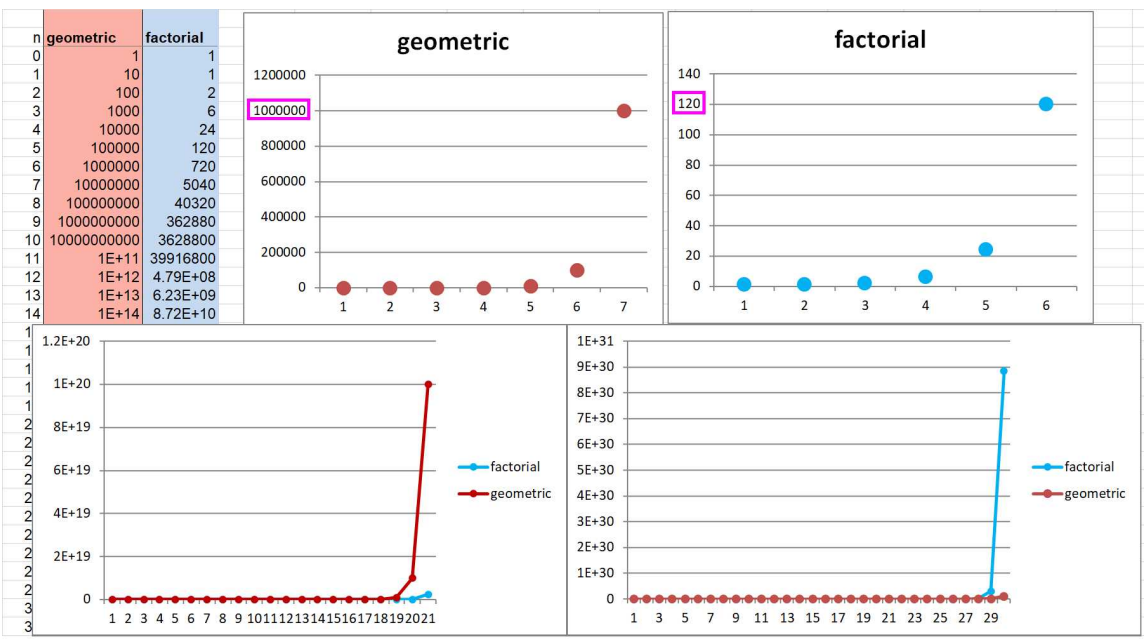
$$n!$$

It reads "n-factorial".

For example, this is how a few initial terms are computed. We progress from left to right in the bottom row by following the arrow $\nearrow$ to find and multiply by the current value of n, then placing the result where the next arrow $\downarrow$ points:

$$\begin{array}{r|ccccccccc} n & 0 & & 1 & & 2 & & 3 & & 4 & \ldots \\ & & \nearrow & \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & \ldots \\ a_n = n! & 1 & & 1 & & 2 & & 6 & & 24 & \ldots \end{array}$$

The factorial exhibits a very fast growth, eventually:



You can see how the factorial (blue) stays behind the geometric progression with ratio $r = 10$ (red) but then leaps ahead.

The factorial appears frequently in calculus and elsewhere. It suffices to point out for now that it counts the number of ways one can *permute* objects.

Example 1.3.22: permutations

Suppose we have 5 guests to be placed around the table. There are 5 seats:

$$\begin{array}{ccc} \text{guests:} & & \text{seats:} \\ A,\ B,\ C,\ D,\ E & \longrightarrow & 1,\ 2,\ 3,\ 4, 5 \end{array}$$

In how many ways can we do this?

We start with the first seat. There are 5 guests to choose from. Once a choice is made, we have 4 guests left.

Therefore, for each of the 5 choices we had, we have 4 guests to be placed at the second seat. The total number is $5 \cdot 4 = 20$ choices. Once a choice is made, we have 3 guests left.

Therefore, for each of the 20 choices we had, we have 3 guests to be placed at the third seat. The total number is $20 \cdot 3 = 60$ choices. Once a choice is made, we have 2 guests left.

Therefore, for each of the 60 choices we had, we have 2 guests to be placed at the fourth seat. The total number is $60 \cdot 2 = 120$ choices.

There is only one guest left and it is placed at the fifth seat.

So, the total number of choices is

$$5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120\,.$$

In general we have to place n objects into n slots, one by one:

- The first object has n options.
- The second $n-1$ options.
- The third $n-2$ options.
- ...
- The last has 1 option the left.

Since the choices are independent of each other, the total number of such placements is

$$n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 = n!$$

Theorem 1.3.23: Number of Permutations

The number of all possible permutations of n objects is equal to n-factorial.

Exercise 1.3.24

In how many ways can you arrange 9 players at the 9 positions on the baseball field?

Warning!

The expression $n!$ is just an abbreviation of the recursive definition of the factorial, not its nth-term formula.

1.4. Repeated addition and repeated multiplication

In this section, we will take a better look at the arithmetic and geometric progressions, side by side.

Remember this simple algebra:

▶ Repeated addition is *multiplication*: $2+2+2 = 2 \cdot 3$.

One can say that that's how multiplication was "invented" – as repeated addition. Next:

▶ Repeated multiplication is *power*: $2 \cdot 2 \cdot 2 = 2^3$.

In either case, we choose to use an *abbreviated* notation for repetitive operations.

Warning!

Some calculators and some computer programs will give you: `-1^2=1` .

By adopting this notation, we create – in addition to the recursive formulas – the nth-term formulas for these sequences, as follows.

Theorem 1.4.1: Formulas for Arithmetic and Geometric Progressions

1. *The nth-term formula for an arithmetic progression with increment a (that starts with a) is*
$$a_n = a \cdot n\,, \quad n = 1, 2, 3, ...$$
2. *The nth-term formula for a geometric progression with ratio a (that starts with a) is*
$$a_n = a^n\,, \quad n = 1, 2, 3, ...$$

So, we face two analogous *conventions* of algebra:

- A real number a that appears n times to be added is replaced with $a \cdot n$.
- A real number a that appears n times to be multiplied is replaced with a^n.

Let's recall the notation and the terminology for the latter:

Base and exponent

$$\overset{\text{base}}{\uparrow}\ a^{\ \downarrow\ n}\ \text{exponent}$$

We will pursue this convenient analogy further and re-discover some familiar algebraic properties. Addition or multiplication, we will ask the same question:

▶ How many times does a appear?

The first set of properties is about the algebraic operations carried out with the values of two sequences. The repetitions are carried out in parallel and then combined together. So, a appears n times, then m times, with a total of $n + m$ times:

Analogy: repeated addition vs. repeated multiplication

$n = 1, 2, 3, ...$	repeated addition	$=$ multiplication	repeated multiplication	$=$ power
Convention:	$\underbrace{a + a + a + ... + a}_{n \text{ times}}$	$= a \cdot n$	$\underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{n \text{ times}}$	$= a^n$
Repeated n times,	$\underbrace{a + a + a + ... + a}_{n \text{ times}}$	$= a \cdot n$	$\underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{n \text{ times}}$	$= a^n$
then m times more.	$+\underbrace{a + a + a + ... + a}_{m \text{ times}}$	$= a \cdot m$	$\cdot\underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{m \text{ times}}$	$= a^m$
Count:	$= \underbrace{a + a + a + ... + a}_{n+m \text{ times}}$		$= \underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{n+m \text{ times}}$	
Property 1:	$a \cdot (n + m)$	$= a \cdot n + a \cdot m$	a^{n+m}	$= a^n \cdot a^m$

This property for addition

$$a \cdot (n + m) = a \cdot n + a \cdot m$$

is called the *Distributive Property*. It "distributes" multiplication over addition and thereby undoes the effect of factoring. The other formula is its analogue:

Theorem 1.4.2: Addition-Multiplication Rule of Exponents

For every $a > 0$ and any natural numbers n and m, we have:

$$a^{n+m} = a^n \cdot a^m$$

We can also see how this formula turns addition into multiplication.

Every property/rule that we discover will often operate as a *shortcut*.

Example 1.4.3: formulas are shortcuts

We use the formula to *expand* the expression on the left:

$$2^{3+2} = 2^3 \cdot 2^2 .$$

Or going backward, we use the formula to *contract* expressions:

$$2^3 \cdot 2^2 = 2^{3+2} = 2^5 .$$

Warning!

We don't "distribute" exponentiation over addition: $a^{n+m} \neq a^n + a^m$.

The second set of properties is also about the algebraic operations carried out – in parallel – with the values of two sequences. This time they have different bases. So, a appears n times, and b appears n times, so $a\&b$ appears n times too:

Analogy: repeated addition vs. repeated multiplication

$n = 1, 2, 3, ...$	repeated addition	= multiplication	repeated multiplication	= power
Convention:	$\underbrace{a + a + a + ... + a}_{n \text{ times}}$	$= a \cdot n$	$\underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{n \text{ times}}$	$= a^n$
Repeated n times.	$\underbrace{a + a + a + ... + a}_{n \text{ times}}$	$= a \cdot n$	$\underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{n \text{ times}}$	$= a^n$
Repeated n times.	$+\underbrace{b + b + b + ... + b}_{n \text{ times}}$	$= b \cdot n$	$\cdot\underbrace{b \cdot b \cdot b \cdot ... \cdot b}_{n \text{ times}}$	$= b^n$
Count:	$= \underbrace{(a + b) + ... + (a + b)}_{n \text{ times}}$		$= \underbrace{(a \cdot b) \cdot ... \cdot (a \cdot b)}_{n \text{ times}}$	
Property 2:	$(a + b) \cdot n$	$= a \cdot n + b \cdot n$	$(a \cdot b)^n$	$= a^n \cdot b^n$

This property for addition

$$(a + b) \cdot n = a \cdot n + b \cdot n$$

is, once again, the *Distributive Property*. It "distributes" multiplication over addition. The analogous property for multiplication "distributes" exponentiation over multiplication:

Theorem 1.4.4: Distributive Rule of Exponents

For every $a, b > 0$ and every natural n, we have:

$$(a \cdot b)^n = a^n \cdot b^n$$

Example 1.4.5: formulas are shortcuts

We use the formula to *expand* the expression on the left:

$$(2 \cdot 3)^5 = 2^5 \cdot 3^5 \,.$$

Or going backward, we can also use the formula to *contract* expressions:

$$2^5 \cdot 3^5 = (2 \cdot 3)^5 = 6^5 \,.$$

Warning!

We also can't "distribute" exponentiation over addition this way: $(a+b)^n \neq a^n + b^n$.

In contrast to the properties above, the next set is about consecutive computations ("compositions"). We repeat the repeated. So, in each row a appears n times, and there are m rows in the table, so a appears a total of $n \cdot m$ times:

Analogy: repeated addition vs. repeated multiplication

$n = 1, 2, 3, ...$		repeated addition	$=$ multiplication	repeated multiplication	$=$ power
Convention:		$\underbrace{a+a+a+...+a}_{n \text{ times}}$	$= a \cdot n$	$\underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{n \text{ times}}$	$= a^n$
Repeated n times,	1.	$\underbrace{a+a+a+...+a}_{n \text{ times}}$	$= a \cdot n$	$\underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{n \text{ times}}$	$= a^n$
m times.	2.	$+\underbrace{a+a+a+...+a}_{n \text{ times}}$	$= a \cdot n$	$\cdot\underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{n \text{ times}}$	$= a^n$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$m.$	$+\underbrace{a+a+a+...+a}_{n \text{ times}}$	$= a \cdot n$	$\cdot\underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{n \text{ times}}$	$= a^n$
Count:		$\underbrace{\qquad\qquad}_{nm \text{ times}}$	$\underbrace{\quad}_{m \text{ times}}$	$\underbrace{\qquad\qquad}_{nm \text{ times}}$	$\underbrace{\quad}_{m \text{ times}}$
Property 3:		$a \cdot (n \cdot m)$	$= (a \cdot n) \cdot m$	$a^{(n \cdot m)}$	$= (a^n)^m$

This property for addition

$$a \cdot (n \cdot m) = (a \cdot n) \cdot m$$

is called the *Associativity Property* of multiplication. It means that multiplications can be re-grouped arbitrarily: the middle number can be "associated" with the last one or the next one. The other formula is its analogue:

Theorem 1.4.6: Multiplication-Exponentiation Rule of Exponents

For every real $a > 0$ and every natural numbers n and m, we have:

$$a^{n \cdot m} = \left(a^n\right)^m$$

We can also see how this formula turns multiplication into "exponentiation".

Example 1.4.7: formulas are shortcuts

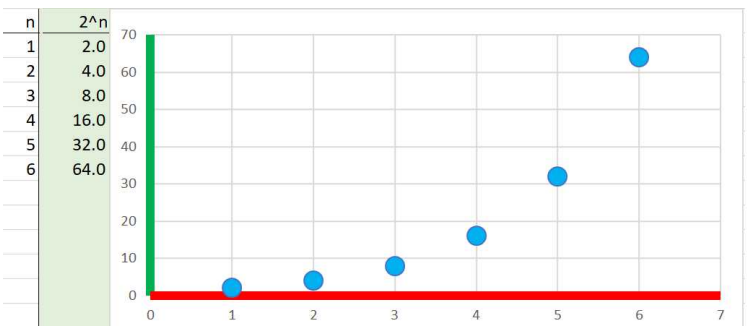
We use the formula to *expand* the expression on the left:

$$2^{3\cdot 4} = (2^3)^4 .$$

Or going backward, we use the formula to *contract* expressions:

$$(2^3)^4 = 2^{3\cdot 4} = 2^{12} .$$

So far, we are facing nothing but a *geometric progression* with ratio a and $n = 1, 2, 3, ...$:



What about $n = 0$? There seems to be a place for it near the origin. But what would be the outcome – and the meaning – of repeating an algebraic operation *zero times*?

We know what it is for addition; we choose:

$$a \cdot 0 = 0$$

In other words, a added to itself 0 times is 0. That's just another *convention*! We adopt one for multiplication too; we choose as follows once and for all.

Definition 1.4.8: exponent equal to zero

The 0th power of any number is 1:

$$a^0 = 1$$

But why? Why not 0? Why not any other number? Because we want the three properties still to be valid! This way we can continue to use them as if nothing has changed.

Theorem 1.4.9: Zero Exponent Rule

The rules of exponents hold when $a^0 = 1$ but fail for any other choice of a^0.

Proof.

Let's check. We plug in $n = 0$ or $m = 0$ and use our convention:

Property 1:	a^{n+m}	$= a^n \cdot a^m$	$n = 0$	$\Longrightarrow$	a^{0+m}	$= a^0 \cdot a^m$	$\Longleftrightarrow$	$a^m = 1 \cdot a^m$ TRUE
Property 2:	$a^n b^n$	$= (ab)^n$	$n = 0$	$\Longrightarrow$	$a^0 b^0$	$= (ab)^0$	$\Longleftrightarrow$	$1 \cdot 1 = 1$ TRUE
Property 3:	a^{nm}	$= (a^n)^m$	$n = 0$	$\Longrightarrow$	a^{0m}	$= (a^0)^m$	$\Longleftrightarrow$	$a^0 = 1^m$ TRUE
			$m = 0$	$\Longrightarrow$	a^{n0}	$= (a^n)^0$	$\Longleftrightarrow$	$a^0 = 1$ TRUE

Exercise 1.4.10

Provide the rest of the proof of the theorem.

Exercise 1.4.11

State the theorem as an equivalence.

Exercise 1.4.12

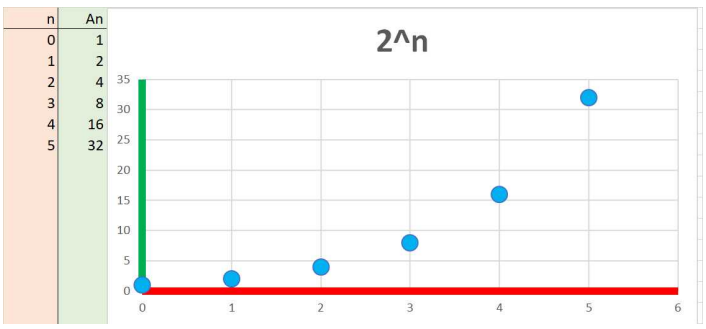
Simplify: (a) $5^0 \cdot 5^3$, (b) $(4 \cdot 3)^2$, (c) $(3^3)^3$, (d) 1^{3+3}, (e) $5^1 \cdot 3^1$, (f) $2^{2 \cdot 2}$. Make up a few of your own. Repeat.

The three properties are still satisfied, and we will continue to use them; choosing anything but $a^0 = 1$ would have ruined them.

From now on, the formula for *geometric progression*

$$a_n = a^n$$

can start with an index value $n = 0$ and a zeroth term a_0:



Exercise 1.4.13

Explain the meaning of a_0 in the compounded interest sequence.

Below is the summary of the algebra of exponents (where a and b *any* are real numbers):

Analogy: repeated addition vs. repeated multiplication

	Multiplication:	**Exponentiation:**
$n = 1, 2, 3, \ldots$	$\underbrace{a + a + a + \ldots + a}_{n \text{ times}} = a \cdot n$	$\underbrace{a \cdot a \cdot a \cdot \ldots \cdot a}_{n \text{ times}} = a^n$
$n = 0$	$a \cdot 0 = 0$	$a^0 = 1$
Rules: 1.	$a \cdot (n + m) = a \cdot n + a \cdot m$	$a^{n+m} = a^n \cdot a^m$
2.	$(a + b) \cdot n = a \cdot n + b \cdot n$	$(a \cdot b)^n = a^n \cdot b^n$
3.	$a \cdot (n \cdot m) = (a \cdot n) \cdot m$	$a^{n \cdot m} = (a^n)^m$

Note that the right-hand sides of the rules are matched: Just replace "+" with "·" and "·" with "^" in the first column and you get those in the second. For example, this is how rule 1 works:

$$\begin{array}{ccccccccc} a & \cdot & (n+m) & = & a & \cdot & n & + & a & \cdot & m \\ a & \wedge & (n+m) & = & a & \wedge & n & \cdot & a & \wedge & m \end{array}$$

The two rules are truly *parallel*. However, this is not the case – warning! – for the left-hand sides of the rules. The reason is that some of the algebra in the left-hand sides has come from *counting* the repetitions, identically for both columns.

So, we use the formulas to "simplify" expressions. Depending on the circumstances, it might mean to expand or it might mean to contract.

We will further continue to expand the idea of exponent in Chapter 4.

A related *convention* is the following:

Definition 1.4.14: factorial of zero

The *factorial of zero* is 1:

$$0! = 1$$

Exercise 1.4.15

Devise notation for repeated exponentiation.

1.5. How to find nth-term formulas for sequences

Let's review how the algebra discussed in the last section allows us to produce a couple of explicit formulas.

First, for an arithmetic progression, we have a recursive formula:

$$a_{n+1} = a_n + b\,,$$

which, written explicitly, takes this form:

$$a_n = a_0 + \underbrace{b + b + b + \ldots + b}_{n \text{ times}}\,.$$

Finally, we have another representation now, without "..."!

Theorem 1.5.1: Formula for Arithmetic Progression

The nth-term formula for an arithmetic progression with increment b and initial term a_0 is the following:

$$a_n = a_0 + b \cdot n, \quad n = 0, 1, 2, 3, \ldots$$

Second, for a geometric progression, we have a formula:

$$a_{n+1} = a_n \cdot r\,,$$

which, written explicitly, takes this form:

$$a_n = a_0 \cdot \underbrace{r \cdot r \cdot r \cdot \ldots \cdot r}_{n \text{ times}}\,.$$

Finally, we have another representation now, without "..."!

Theorem 1.5.2: Formula for Geometric Progression

The nth-term formula for a geometric progression with ratio r and initial term a_0 is the following:

$$a_n = a_0 \cdot r^n, \quad n = 0, 1, 2, 3, \ldots$$

These two sequences, given originally by their *recursive formulas*, are now presented by their nth-term formulas.

It is also sometimes possible, starting with a *list*, to work your way backwards and invent such a formula.

Example 1.5.3: finding the nth term

What is the formula for this sequence:

$$1,\ 1/2,\ 1/4,\ 1/8,\ ...?$$

First, we notice that these are all fractions and their numerators are just 1's:

$$\frac{1}{1},\ \frac{1}{2},\ \frac{1}{4},\ \frac{1}{8},\cdots$$

Second, the denominator is multiplied by 2 every time. It is the same as multiplying the whole fraction by $\frac{1}{2}$. The recursive formula is then:

$$a_{n+1} = a_n \cdot \frac{1}{2}\,.$$

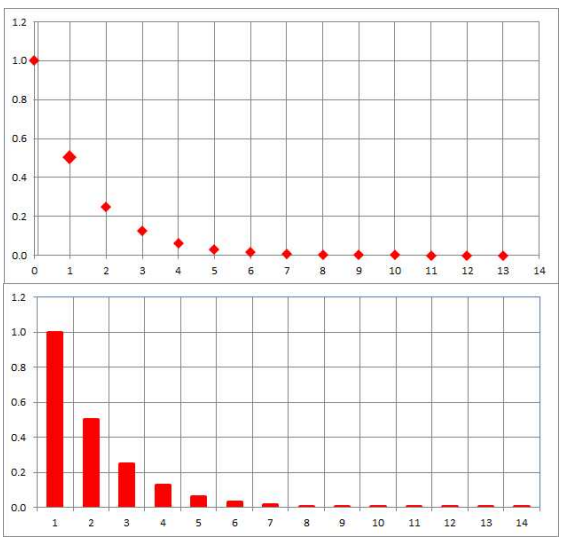
It is a geometric progression! Its ratio is $r = 1/2$ and its initial term is $a_0 = 1$. According to the last theorem, we have

$$a_n = 1 \cdot \left(\frac{1}{2}\right)^n = \frac{1}{2^n}\,.$$

It is possible to see the pattern from the beginning: The denominators are the powers of 2, i.e.,

$$a_0 = 1,\ a_1 = \frac{1}{2},\ a_2 = \frac{1}{2^2},\ a_3 = \frac{1}{2^3},\ ...$$

With this formula, we can plot more terms:



Exercise 1.5.4

What is the formula for the sequence if we require it to start with $a_1 = 3$ instead?

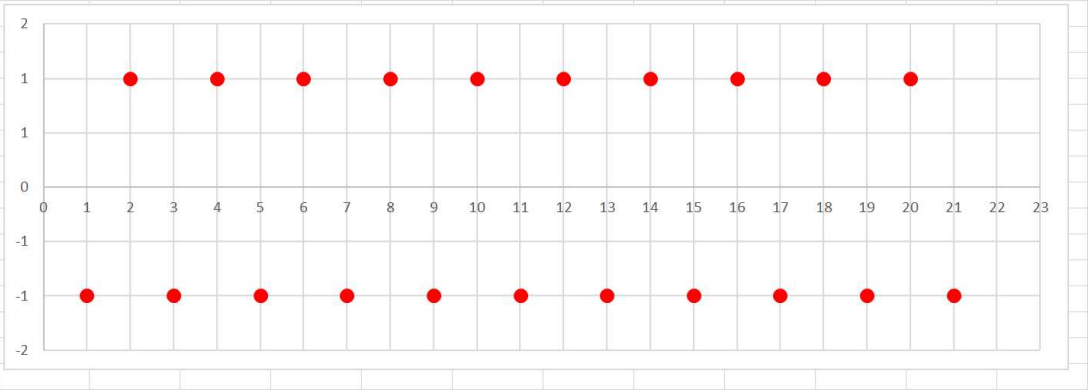
Example 1.5.5: alternating sequence

What is the formula for this sequence:

$$1,\ -1,\ 1,\ -1,\ ...?$$

The plot is simple:

n	An
1	-1
2	1
3	-1
4	1
5	-1
6	1
7	-1
8	1
9	-1
10	1
11	-1
12	1
13	-1
14	1



First, we notice these numbers are just all 1's and only the sign *alternates*. We write it in a more convenient form:

$$a_0 = 1,\ a_1 = -1,\ a_2 = 1,\ a_3 = -1,\ ...$$

The pattern is clear and the formula can be written for the two cases separately:

$$a_n = \begin{cases} -1 & \text{if } n \text{ is even,} \\ 1 & \text{if } n \text{ is odd.} \end{cases}$$

This qualifies as its nth-term formula but there is a more compact version:

$$a_n = (-1)^{n+1}\,.$$

We were able to get rid of "..."!

The observation is worth recording.

Theorem 1.5.6: Formula for Alternating Sequence

The nth-term formula for the alternating sequence,

$$a_0 = 1,\ a_1 = -1,\ a_2 = 1,\ a_3 = -1,\ ...,$$

is given by the following:

$$a_n = (-1)^n\,, \quad n = 0, 1, 2, 3, ...$$

Exercise 1.5.7

What is the formula for the sequence if it starts with $a_1 = 1$ instead?

Exercise 1.5.8

The alternating sequence above is a _____ sequence.

Exercise 1.5.9

Point out a pattern in each of the following sequences and suggest a formula for its nth term whenever possible:

1. 1, 3, 5, 7, 9, 11, 13, 15, ...
2. .9, .99, .999, .9999, ...
3. 1/2, −1/4, 1/8, −1/16, ...
4. 1, 1/2, 1/3, 1/4, ...
5. 1, 1/2, 1/4 , 1/8, ...
6. 2, 3, 5, 7, 11, 13, 17, ...

7. 1, −4, 9, −16, 25, ...
8. 3, 1, 4, 1, 5, 1, 9, ...

Example 1.5.10: regular deposits, results computed

A person starts to deposit $20 every month into his bank account that already contains $1000:

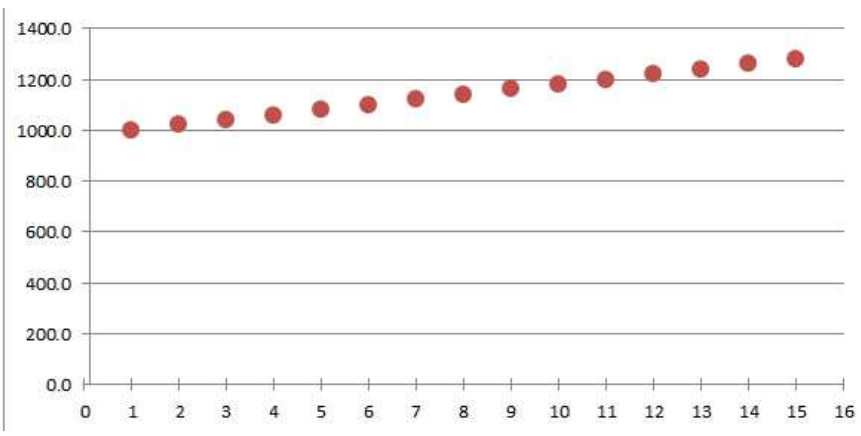
$$a_{n+1} = a_n + 20\,.$$

We have now a formula for the nth term:

$$a_n = 1000 + 20 \cdot n\,,$$

assuming that $a_0 = 1000$:

time	amount
1	1000.00
2	1020.00
3	1040.00
4	1060.00
5	1080.00
6	1100.00
7	1120.00
8	1140.00
9	1160.00
10	1180.00
11	1200.00
12	1220.00
13	1240.00
14	1260.00
15	1280.00



Exercise 1.5.11

How long will it take to double your money?

Example 1.5.12: compounded interest, results computed

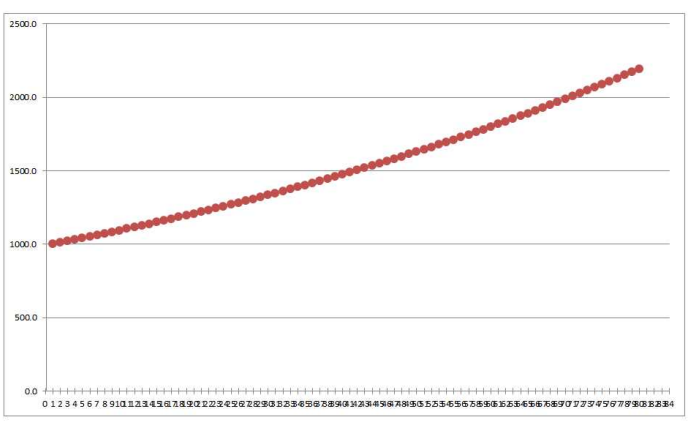
A person deposits $1000 into his bank account that pays 1% interest every year:

$$a_{n+1} = a_n \cdot 1.01\,.$$

We have now a formula:

$$a_{n+1} = 1000 \cdot 1.01^n\,.$$

This is the graph:



Here is the general setup. A person deposits a_0 dollars to his bank account. Suppose the account pays R, the decimal representation of the APR compounded annually, i.e., .10 for 10 percent etc. The total amount is then multiplied by $1 + R$ at the end of each year. Now, if a_n stands for the amount in the bank account after n years, we have the following recursive formula:

$$a_{n+1} = a_n \cdot (1 + R)\,.$$

A verbal description of the model – *the growth is proportional to the size of current amount* – reveals that this is a geometric progression. Its nth-term formula is the following:

$$a_n = a_0 \cdot (1+R)^n, \ n = 1, 2, 3, ...$$

Exercise 1.5.13

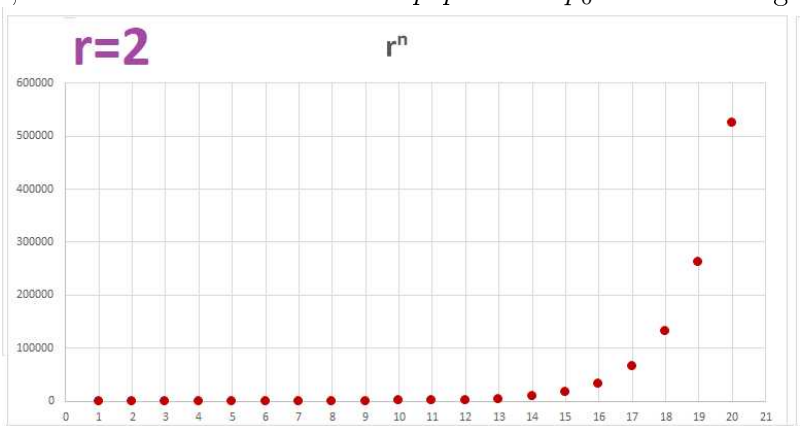
How long will it take to double your money?

Example 1.5.14: bacteria multiplying

Suppose we have a population of bacteria that doubles every day. For example, we can imagine that every one of them divides in half every day. Let p_n be the number of bacteria after n days:

$$\underbrace{p_{n+1}}_{\text{population: at time } n+1} = 2 \ \underbrace{p_n}_{\text{at time } n}$$

To know p_n for all n's, we need to know the *initial population* p_0. This is the graph:



We have a geometric progression. Its nth-term formula is:

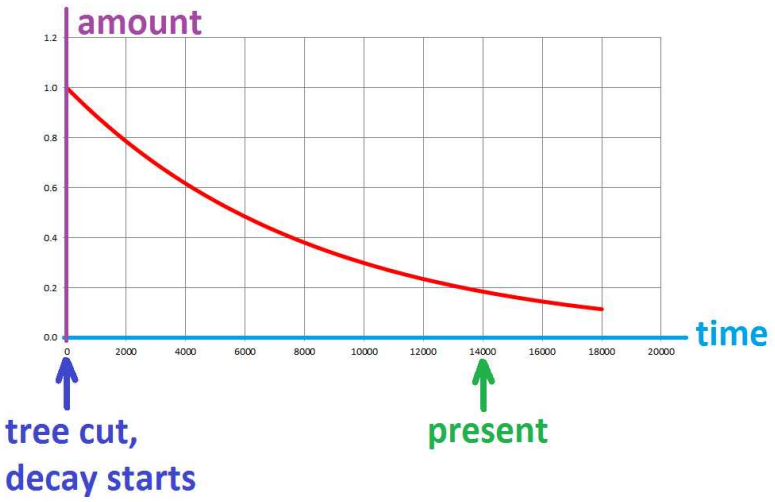
$$p_n = p_0 2^n \, .$$

Example 1.5.15: radioactive decay and radiocarbon dating

It is known that once a tree is cut, the radioactive carbon it contains starts to decay. It loses half of its mass over a period of time of a certain length called the *half-life* of the element. The loss, therefore, follows the familiar exponential decay model:

$$a_{n+1} = a_n \cdot \frac{1}{2} \, .$$

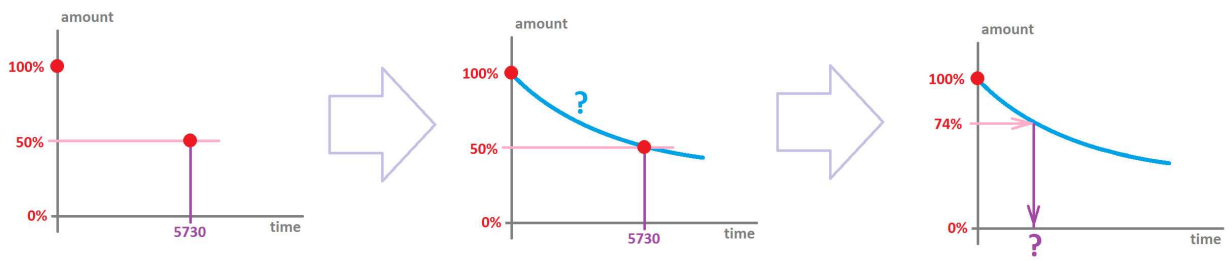
A verbal description of the model – *the decay is proportional to the current amount* – reveals that this is a geometric progression. The difficulty is that here n is *not* the number of years but the number of half-lives! For example, the percentage of this element, ^{14}C, left is plotted below under the assumption that its half-life is 5730 years (i.e., the time it takes to go from 100% to 50%):



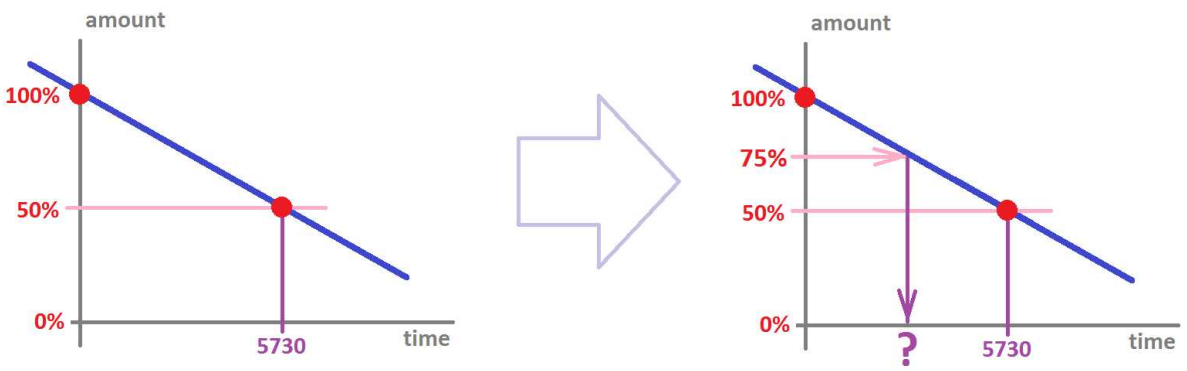
The idea we will develop is as follows:

- Find you the element's half-life.
- Measure the percentage of the element present vs. the amount normally present.
- Calculate the time when tree was cut.

Suppose a parchment has 74% of ^{14}C left. How old is it? Let's take a closer look at the graph:



Unfortunately, the model measures time in multiples of the half-life, 5730 years. Any period shorter than that is out of reach for now. We can try to estimate the answer by assuming that this is an arithmetic progression, yearly:



The period, therefore, is estimated to be close to:

$$\frac{1}{4}\text{-life} = \frac{1}{2}\left(\frac{1}{2}\text{-life}\right) = 2865\,.$$

In Chapter 4, we will learn how to deal with fractional time periods and solve the problem exactly.

Example 1.5.16: Newton's Law of Cooling

A cooler object in a warmer environment heats up, and a warmer objects cools down in a colder environment.

How fast that happens depends on the difference between the object's temperature T_n (at time n) and the room temperature R. This number determines the new temperature T_{n+1}. The law states that

the *rate of cooling* is proportional to this difference:

$$T_{n+1} - R = (T_n - R) \cdot k\,,$$

for some $k < 1$. We, therefore, have a recursive formula:

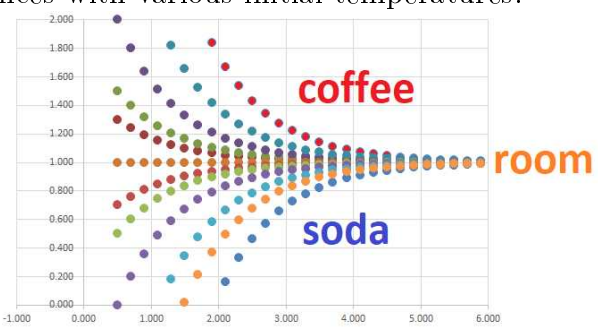
$$T_{n+1} = R + (T_n - R) \cdot k\,.$$

Unless $R = 0$, this is *not* a geometric progression (but $T_n - R$ is).

There are three cases determined by the initial temperature:

- $T_0 > R$: cooling,
- $T_0 = R$: unchanging,
- $T_0 < R$: warming.

We plot below several sequences with various initial temperatures:



Exercise 1.5.17

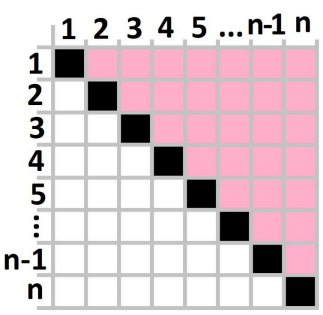
Find the nth-term formula for the sequence in the above example.

Collectively, these examples are called *exponential models*; they are characterized by the dynamics that comes from repeated multiplication by the same number.

There are many other types of models.

Example 1.5.18: number of plays in round robin

If we have n teams to play each other exactly once, how many games do we have to plan for? A table commonly used for such a tournament is below:



The table reveals the following: The first team is to play $n-1$ games. Likewise, the second also is to play $n-1$ games but one less is actually counted as it is already on the first list. The third is to play $n-1$ games but two less is actually counted as they are already on the first and second lists. And so on. The total is

$$(n-1) + (n-2) + \ldots + 2 + 1\,.$$

We can treat these $n-1$ numbers as a recursive sequence:

$$a_1 = 1,\ a_{n+1} = a_n + n\,.$$

How do we find an explicit, direct formula for the nth term of this sequence?

The table suggests the answer. The total number of cells in an $n \times n$ table is n^2. Without the ones on the diagonal, it's $n^2 - n$. Finally, we take only half of those: $(n^2 - n)/2$.

As a purely mathematical conclusion, the sum of n consecutive integers starting from 1 is the following:

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}\,.$$

We were able to get rid of "..."! (The formula will find another use in Chapter 3IC-1.)

Exercise 1.5.19

Show that the sum of consecutive *odd* integers satisfies the following:

$$1 + 3 + 5 + 7 + \ldots + (2n - 1) = n^2\,.$$

Can we always find an explicit formula? No. We have already seen an example of a recursively defined sequence without nth-term formula, the factorial, $n!$.

Example 1.5.20: deposits and interest, together

From earlier, we know that if we deposit money into our bank account and receive interest, the recursive formula is simple:

$$a_{n+1} = a_n \cdot 1.05 + 200\,.$$

But is there a *direct* nth-term formula? It's too cumbersome to be of any use:

$$a_n = \Big(\ldots\big((a_0 \cdot 1.05 + 200) \cdot 1.05 + 200\big)\ldots{}_{\text{repeat n times}}\Big) \cdot 1.05 + 200\,.$$

Since we don't know how to get rid of "...", we are left with a formula which is just the recursive definition of the sequence in disguise. The best we can do is to say that the sequence is increasing.

Example 1.5.21: restricted growth

Suppose our bacteria live in a *jar*. We are then forced to add another effect to our population model:

1. The rate of reproduction will still be proportional to the current population – but only when the population size is "small" and the effects of the restricted room are negligible; and
2. Once it gets "large", starvation starts – the growth rate will decrease at a rate proportional to how close the population is to the theoretical *carrying capacity*.

For example, let's suppose, for our bacteria the jar can accommodate 1000. Then the recursive formula $p_{n+1} = 2p_n$ is modified by adding a new factor to accommodate the second feature of the model:

$$p_{n+1} = 2p_n \cdot \frac{1000 - p_n}{1000}\,.$$

The new factor indicates how close we are, proportionally, to this limit.

Let's take a specific example. If the current population is $p_n = 800$, the next is

$$p_{n+1} = 2 \cdot 800 \cdot \frac{1000 - 800}{1000} = 320\,.$$

It goes *down* because it is too close to the capacity. But the next one,

$$p_{n+2} = 2 \cdot 320 \cdot \frac{1000 - 320}{1000} = 435.2\,,$$

is *up* again! It might continue to go up and down.

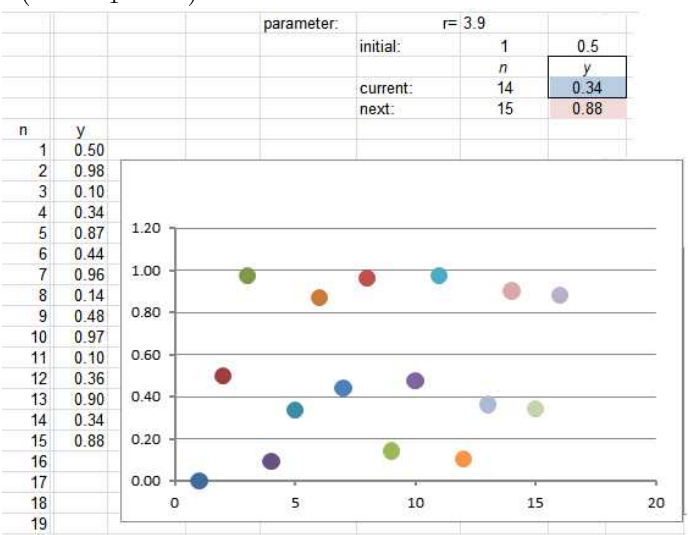
In general, we define a sequence to represent the *proportion* of the largest possible population as follows:

$$a_{n+1} = ra_n(1 - a_n)\,,$$

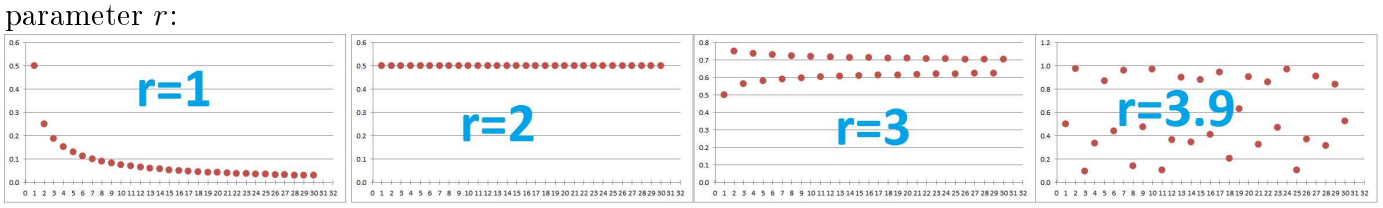
where $r > 0$ is a parameter. So, the first factor tries to, say, double the population, while the second holds it back. For the spreadsheet, the formula refers to the previous state, `R[-1]C`:

```
=R2C2*R[-1]C*(1-R[-1]C)
```

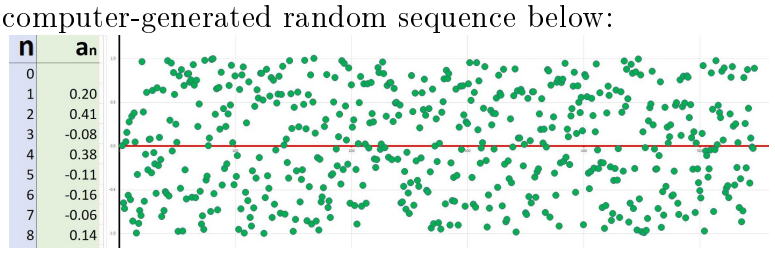
while `R2C2` is the cell that contains the value of the parameter r. We present a computation with a specific parameter $r = 3.9$ (and $a_1 = .5$):



Such a sequence is called a *logistic sequence*. Its dynamics dramatically depends on the choice of the parameter r:



Above, the sequence starts with 1/2 and then shows several different types of dynamics: decay, equilibrium, oscillating convergence to equilibrium, and chaos. There is no nth-term formula. The chaotic one looks similar to a computer-generated random sequence below:



Exercise 1.5.22

What if new bacteria are introduced to the jar at a constant rate?

1.6. The algebra of exponents

Let's review our analysis from earlier where the exponents – by analogy with products – come from:

Analogy: repeated addition vs. repeated multiplication

	Multiplication:	**Exponentiation:**
$n = 1, 2, 3, \ldots$	$\underbrace{a + a + a + \ldots + a}_{n \text{ times}} = a \cdot n$	$\underbrace{a \cdot a \cdot a \cdot \ldots \cdot a}_{n \text{ times}} = a^n$
$n = 0$	$a \cdot 0 = 0$	$a^0 = 1$
Rules: 1.	$a(n+m) = a \cdot n + a \cdot m$	$a^{n+m} = a^n \cdot a^m$
2.	$(a+b)n = a \cdot n + b \cdot n$	$(ab)^n = a^n b^n$
3.	$a \cdot (nm) = (a \cdot n) \cdot m$	$a^{nm} = (a^n)^m$

Exercise 1.6.1

Represent as a power of 3: (a) $3^3 \cdot 9$, (b) $(9 \cdot 27)^5$.

This new algebra helps us to understand *factoring* of integers. For example, the last factorization below is best:

$$20 = 4 \cdot 5 = 2 \cdot 2 \cdot 5 = 2^2 5^1 .$$

But we can improve it even further:

$$20 = 2^2 3^0 5^1 .$$

Let's review what we know about factoring:

Theorem 1.6.2: Rules of Factoring Integers

- *A prime number cannot be factored into smaller integers* > 1*:*

$$2, 3, 5, 7, 11, 13, \ldots$$

- ***Fundamental Theorem of Arithmetic:*** *Every integer can be represented as the product of primes.*
- ***Fundamental Theorem of Arithmetic:*** *This representation is unique, up to the order of the factors:*

$$q = 2^m \cdot 3^n \cdot 5^k \cdot \ldots$$

- *The multiplicity of a prime factor is how many times it appears in the factorization.*

Then, in the prime factorization

$$20 = 2^2 3^0 5^1 ,$$

the multiplicity of 2 is 2, of 5 is 1, and the rest are 0s. And there are no other factorizations!

Example 1.6.3: prime factorization

Let's find *the* prime factorization of 1176.

Our strategy is to divide by larger and larger primes, switching to the next one when we can't. We start with 2 and then go up. We try to divide by 2 and discover that it is possible:

$$1176/2 = 588 .$$

We start the list of factors with this 2. We will continue to try to divide by 2. It is possible:

$$588/2 = 294 ,$$

and we add this 2 to our list of factors: $2, 2$. We continue to divide by 2:

$$294/2 = 147\,.$$

It works and we have: $2, 2, 2$. But the result is odd! We, therefore, move on to the next prime, 3:

$$147/3 = 49\,.$$

It works and our list becomes: $2, 2, 2, 3$. The result is not divisible by 3 and we, therefore, move on to the next prime, 5. We fail to divide by 5. The next prime works:

$$49/7 = 7\,.$$

This last number is also prime and we are done with our list:

$$2, 2, 2, 3, 7, 7\,.$$

The prime factorization is as follows:

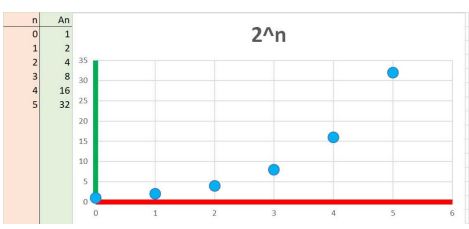
$$1176 = 2 \cdot 2 \cdot 2 \cdot 3 \cdot 7 \cdot 7 = 2^3 3^1 5^0 7^2\,.$$

Here, 3, 1, 0, 2 are the respective multiplicities.

Exercise 1.6.4

Factorize each into the product of powers of primes: (a) 500, (b) 660, (c) 1024.

We are familiar with *geometric progressions*. They start with index $n = 0$:



However, there are other values of n that might also interest us!

Example 1.6.5: bacteria multiplying, the past

Suppose our bacteria double in number every day and the current population is 1000. Then, tomorrow it will be $1,000 \cdot 2$, and so on. After n days, it's

$$1,000 \cdot 2^n\,.$$

But how many were there *yesterday*? Half of it, 500! Does this number fit into our geometric progression a_n? Can we choose $n = -1$? Let's try:

$$a_{-1} = 500 = 1,000/2 \stackrel{?}{=} 1,000 \cdot 2^{-1}\,.$$

It seems that if we go into the future, we multiply, and if we go to the past, we'll need to *divide*.

Just as earlier in this chapter we face new circumstances and, once again, we ask ourselves: Can we proceed *without changing the rules*?

Can the exponents be *negative*?

We start with $n = -1$. Can -1 be the exponent, such as 2^{-1}? And what would be the outcome – and the meaning – of repeating an algebraic operation *minus* 1 *times*?!

We will need another *convention*! We choose, for $a \neq 0$, the following *notation*:

$$a^{-1} = \frac{1}{a}\,.$$

The choice is dictated by our desire for the three properties to be satisfied!

Theorem 1.6.6: Exponent Equal to -1

The rules of exponents hold for $a^{-1} = 1/a$ but fail for any other choice of a^{-1}.

Proof.

Let's check. We plug in $n = \pm 1$ and $m = \pm 1$, use our convention, and then apply one of the properties presented above:

1:	a^{n+m}	$= a^n \cdot a^m$	$n=-1,\ m=1$	$\Longrightarrow$	a^{-1+1}	$= a^{-1}\cdot a^1$	$\Longleftrightarrow$	$a^0 = \frac{1}{a}\cdot a$	TRUE
2:	$a^n b^n$	$= (ab)^n$	$n=-1$	$\Longrightarrow$	$a^{-1}b^{-1}$	$= (ab)^{-1}$	$\Longleftrightarrow$	$\frac{1}{a}\cdot\frac{1}{b} = \frac{1}{ab}$	TRUE
3:	a^{nm}	$= (a^n)^m$	$n=-1,\ m=1$	$\Longrightarrow$	$a^{(-1)1}$	$= (a^{-1})^1$	$\Longleftrightarrow$	$a^{-1} = \frac{1}{a}$	TRUE
			$n=1,\ m=-1$	$\Longrightarrow$	$a^{1(-1)}$	$= (a^1)^{-1}$	$\Longleftrightarrow$	$a^{-1} = a^{-1}$	TRUE

Exercise 1.6.7

Provide the rest of the proof.

So, this is our **convention**: *Multiplying by a^{-1} means dividing by a.*

Exercise 1.6.8

Represent as exponents: (a) $.2$, (b) $.25$, (c) $.0001$.

Now, the rest of the negative numbers.

First, let's take another look at our analogy. While multiplication by a positive integer means repeated addition, multiplication by a *negative* integer means repeated *subtraction* (the "inverse" of addition):

$$a(-n) = -an = 0 - a - a - a - \ldots - a\,.$$

Similarly, while exponentiation by a positive integer means repeated multiplication, exponentiation by a *negative* integer means repeated *division* (the "inverse" of multiplication):

$$a^{-n} = 1 \div a \div a \div a \div \ldots \div a\,.$$

Our convention for any $n = 1, 2, 3, \ldots$ will be, once and for all, the following:

Definition 1.6.9: negative exponent

The power of the negative of a number is the division by the power of that number:

$$a^{-n} = \frac{1}{a^n}$$

To confirm, we observe this:

$$\begin{array}{rcl} \underbrace{a^{-1}\cdot a^{-1}\cdot a^{-1}\cdot \ldots \cdot a^{-1}}_{n \text{ times}} & = & \left(a^{-1}\right)^n \\ || & & || \\ 1\underbrace{\div a \div a \div a \div \ldots \div a}_{n \text{ times}} & = & \left(\frac{1}{a}\right)^n \end{array}$$

Theorem 1.6.10: Negative Exponent Rule

The rules of exponents hold for $a^{-n} = 1/a^n$, where $n > 0$, but fail for any other choice of a^{-n}.

Exercise 1.6.11

Provide the proof of the theorem.

Example 1.6.12: formulas are shortcuts

Once again, we see these properties as *shortcuts*:

$$\frac{2^5}{2^2} = 2^5 \cdot 2^{-2} = 2^{5-2} = 2^3\,.$$

Exercise 1.6.13

Represent as a power of 4: (a) $4^3 \cdot .25$, (b) $.25/4^2$. Represent as a power of 2.

Here is the summary of what we have established:

Analogy: repeated addition vs. repeated multiplication

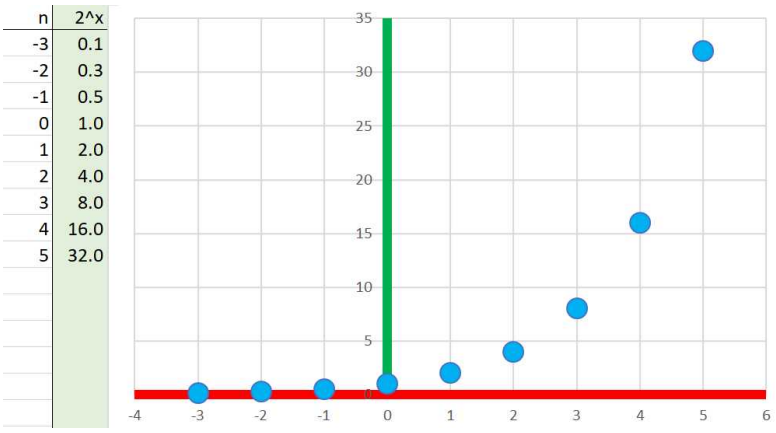
	Multiplication:	**Exponentiation:**
Conventions:		
$n = 1, 2, \ldots$	$\underbrace{a + a + a + \ldots + a}_{n \text{ times}} = a \cdot n$	$\underbrace{a \cdot a \cdot a \cdot \ldots \cdot a}_{n \text{ times}} = a^n$
$n = 0$	$a \cdot 0 = 0$	$a^0 = 1$
$n = -1, -2, \ldots$	$0\underbrace{-a - a - a - \ldots - a}_{n \text{ times}} = a \cdot (-n)$	$1\underbrace{\div a \div a \div a \div \ldots \div a}_{n \text{ times}} = a^{-n}$
	$\underbrace{(-a) + (-a) + \ldots + (-a)}_{n \text{ times}} = (-a) \cdot n$	$\underbrace{\frac{1}{a}\cdot\frac{1}{a}\cdot\ldots\cdot\frac{1}{a}}_{n \text{ times}} = \left(\frac{1}{a}\right)^n$
	$-(\underbrace{a + a + a + \ldots + a}_{n \text{ times}}) = -(a \cdot n)$	$1 \div (\underbrace{a \cdot a \cdot a \cdot \ldots \cdot a}_{n \text{ times}}) = \frac{1}{a^n}$
Rules:		
1.	$a \cdot (n + m) = a \cdot n + a \cdot m$	$a^{n+m} = a^n \cdot a^m$
2.	$(a + b) \cdot n = a \cdot n + b \cdot n$	$(a \cdot b)^n = a^n b^n$
3.	$a \cdot (n \cdot m) = (a \cdot n) \cdot m$	$a^{n \cdot m} = (a^n)^m$

So, the three properties are still satisfied, and we will continue to use them with no regard for a possible negativity of the exponent. There is no need to count anymore.

These are the possible values of n from now on:

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$$

In the graph, we notice that we *multiply* by a as we move right (increase) and *divide* by a as we move left (decrease):



n	2^x
-3	0.1
-2	0.3
-1	0.5
0	1.0
1	2.0
2	4.0
3	8.0
4	16.0
5	32.0

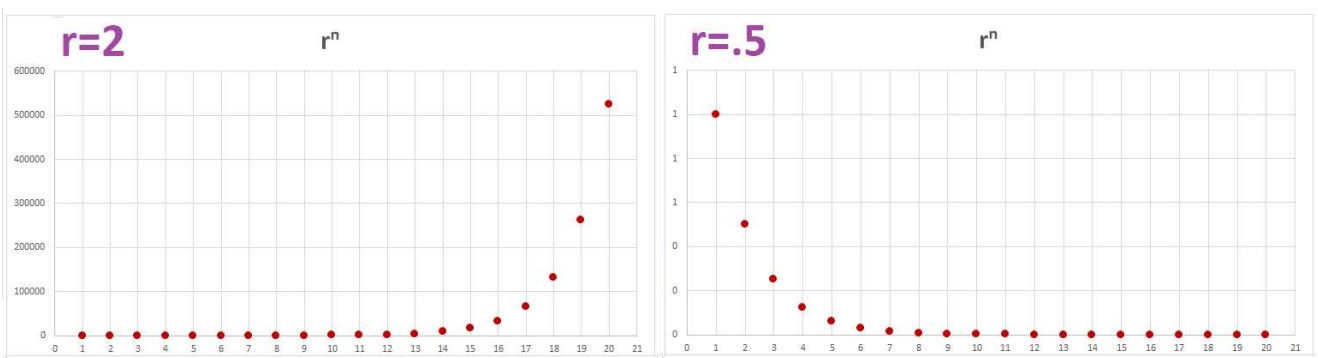
However, if we start at any point and proceed to the right, we only multiply. This is our conclusion.

Theorem 1.6.14: Monotonicity of Geometric Progression

A geometric progression r^n is

- *increasing if $r > 1$, and*
- *decreasing if $0 < r < 1$.*

This is what the graphs look like:



Exercise 1.6.15

Prove the theorem.

Exercise 1.6.16

If our bacteria double in number every day and the current population is 1024, how many were there two days ago?

The integers as the possible moments of time still miss some of the numbers that might interest us in these models! For example, suppose our bacteria double in number every day starting with 1. What happens after 10.5 days? We'll need to figure out the meaning of this:

$$n = 10.5, \quad 2^{10.5} = ?$$

We address the fractional exponents in Chapter 4.

1.7. The Binomial Formula

We know that we aren't supposed to distribute *powers over addition* (an unforgivable mistake!):

$$(a+b)^2 \neq a^2+b^2\,.$$

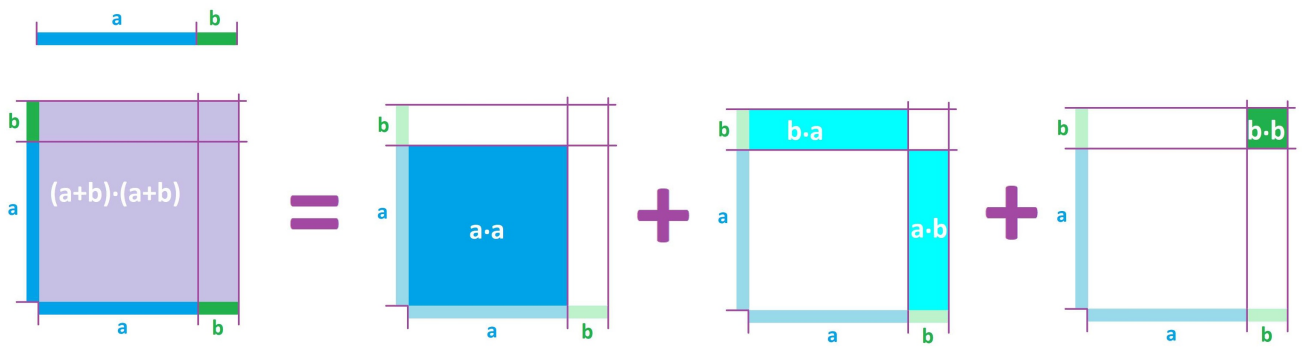
Instead, we find what it is equal to by distributing *multiplication over addition*:

$$\begin{aligned}(a+b)^2 &= (a+b)\cdot(a+b)\\ &= (a+b)\cdot a+(a+b)\cdot b\\ &= (a\cdot a+b\cdot a)+(a\cdot b+b\cdot b)\\ &= a^2+2ab+b^2\,.\end{aligned}$$

Warning!

We repeat the equal sign "=" every time because it stands for the verb "is" and we want our sentences to be grammatically correct.

The meaning of the formula is revealed if we interpret $(a+b)^2$ as the *area of a square* with side $a+b$:



The "end" terms correspond to the two smaller squares, $a\times a$ and $b\times b$, and the "intermediate" terms correspond to the rectangles, $a\times b$ and $b\times a$.

The result is a handy formula:

$$(a+b)^2=a^2+2ab+b^2$$

It is very useful, especially when we need to *factor* an expression; we simply read the formula from right to left:

$$a^2+2ab+b^2=(a+b)^2\,.$$

Example 1.7.1: a complete square

To factor $4x^2+4x+1$, we match it with the left-hand side of the formula above (we assume that the numbers are positive) as follows:

	a^2	$+$	$2ab$	$+$	b^2	$=$	$(a+b)^2$	We match vertically.
	$4x^2$	$+$	$4x$	$+$	1	$=$	$?$	
$\Longrightarrow$	$a^2=4x^2$		$\ldots$		$b^2=1$			The equations that come from the matching.
$\Longrightarrow$	$a=2x$		$\ldots$		$b=1$			The equations are solved.
$\Longrightarrow$	$4x^2$	$+$	$4x$	$+$	1	$=$	$(2x+1)^2$	The middle term also checks out.

We have a complete square:

$$4x^2 + 4x + 1 = (2x + 1)^2 .$$

It is more compact than the original and we can notice certain facts about it that used to be invisible. For example, the expression can't be negative!

Warning!

We do not put the equal sign "=" between equations because each is a separate sentence, and the quantities aren't equal.

Exercise 1.7.2

Factor $9x^2 + 12xy + 4y^2$.

What about the *higher powers*? To get the cubic power and above, we simply continue multiplying by $(a+b)$ and applying the *Distributive Property*:

$$\begin{aligned}
(a+b)^3 &= (a+b)^2 \cdot (a+b) \\
&= (a^2 + 2ab + b^2) \cdot (a+b) \\
&= (a^2 + 2ab + b^2) \cdot a + (a^2 + 2ab + b^2) \cdot b \\
&= (a^2 \cdot a + 2ab \cdot a + b^2 \cdot a) + (a^2 \cdot b + 2ab \cdot b + b^2 \cdot b) \\
&= (a^3 + 2a^2b + ab^2) + (a^2b + 2ab^2 + b^3) \\
&= a^3 + 3a^2b + 3ab^2 + b^3 .
\end{aligned}$$

The result is another handy formula:

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

Example 1.7.3: a complete cube

To factor $x^3 + 3x^2 + 3x + 1$, we match it with the left-hand side of the formula above:

$$\begin{array}{rccccccccc}
 & a^3 & + & 3a^2b & + & 3ab^2 & + & b^3 & = & (a+b)^3 \\
 & x^3 & + & 3x^2 & + & 3x & + & 1 & = & ? \\
\Longrightarrow & a^3 = x^3 & & .. & & .. & & b^3 = 1 & & \\
\Longrightarrow & a = x & & .. & & .. & & b = 1 & & \\
\Longrightarrow & x^3 & + & 3x^2 & + & 3x & + & 1 & = & (x+1)^3
\end{array}$$

The "intermediate" terms also checks out.

Exercise 1.7.4

Interpret $(a+b)^3$ as the volume of a cube with side $a+b$ and illustrate the meaning of the "intermediate" terms in the above formula.

The two-term expression $a + b$ is called a *binomial*. In the formula ($m = 2$ and $m = 3$ above),

$$(a+b)^m = a^m + \ldots + b^m ,$$

we call the left-hand side a *binomial power* and the right-hand side the *binomial expansion*.

Where these terms come from? We pick one term from each binomial and there are a total m of them:

$$(a+b) \cdot (a+b) \cdot \ldots \cdot (a+b) = \quad \text{sum of terms of the type } a^n \cdot b^k .$$

As we arrange the terms according to the decreasing powers of a, a pattern starts to emerge! As the power of a is decreasing, that of b is increasing. Moreover, the *combined power of a and b*, i.e., the sum of the two powers, is $m = 3$, just as it was $m = 2$ in the last formula. Let's make a table for the binomials that starts with $m = 1$:

$m = 1$	$(a+b)^1$	$=$	a^1b^0	$+$	a^0b^1							
powers:	1	$=$	$1+0$	$=$	$0+1$							
$m = 2$	$(a+b)^2$	$=$	a^2b^0	$+$	$2a^1b^1$	$+$	a^0b^2					
powers:	2	$=$	$2+0$	$=$	$1+1$	$=$	$0+2$					
$m = 3$	$(a+b)^3$	$=$	a^3b^0	$+$	$3a^2b^1$	$+$	$3a^1b^2$	$+$	a^0b^3			
powers:	3	$=$	$3+0$	$=$	$2+1$	$=$	$1+2$	$=$	$0+3$			
$m = 4$	$(a+b)^4$	$=$	a^4b^0	$+$	$\boxed{?}a^3b^1$	$+$	$\boxed{?}a^2b^2$	$+$	$\boxed{?}a^1b^3$	$+$	a^0b^4	
powers:	4	$=$	$4+0$	$=$	$3+1$	$=$	$2+2$	$=$	$1+3$	$=$	$0+4$	

We also included – as a guess – the case $m = 4$. We see that $m+1$ is the number of ways we can represent m as a sum of two non-negative numbers. Below we put the results in a spreadsheet with the formula:

```
=RC2+R2C
```

The expansions above follow the diagonal as marked:

		powers of b							
	n	0	1	2	3	4	5	6	7
powers of a	0	0	1	2	3	4	5	6	7
	1	1	2	3	4	5	6	7	8
	2	2	3	4	5	6	7	8	9
	3	3	4	5	6	7	8	9	10
	4	4	5	6	7	8	9	10	11
	5	5	6	7	8	9	10	11	12
	6	6	7	8	9	10	11	12	13
	7	7	8	9	10	11	12	13	14
	8	8	9	10	11	12	13	14	15
	9	9	10	11	12	13	14	15	16

In each, the combined power of a and b remains the same and it grows by 1 as we move to the next diagonal.

The behavior of the powers is clear; it's the *coefficients* that present a challenge. Let's describe the former first.

Theorem 1.7.5: Number of Terms in Binomial Expansion

The expansion of the mth power $(a+b)^m$ of the binomial $(a+b)$ consists of $m+1$ terms of the form a^nb^k, each of which has the sum of powers of a and b equal to m.

Proof.

If all of the terms of the expansion of $(a+b)^m$ have the sum of powers equal to m, then what happens in the next:

$$(a+b)^{m+1} = (a+b)^m \cdot (a+b) \ ?$$

According to the *Distributive Property*, each of the terms of the expansion of $(a+b)^m$ is multiplied by a or by b. Therefore, the total power goes up by 1!

Exercise 1.7.6

Provide the missing parts of the proof.

Once again, we have:

$$(a+b)^m = \text{ sum of terms of the type } a^n \cdot b^k \, .$$

We want to know *how many times* each term appear in the expansion.

In the summary below, we show the powers but the *coefficients* of the terms remain unknown:

What are the coefficients?

$$\begin{array}{rcccccccccccc} (a+b)^m & = & a^m b^0 & + & \boxed{?}a^{m-1}b^1 & + & \boxed{?}a^{m-2}b^2 & +\ldots+ & \boxed{?}a^1 b^{m-1} & + & a^0 b^m \\ m & = & m+0 & = & (m-1)+1 & = & (m-2)+2 & =\ldots= & 1+(m-1) & = & 0+m \end{array}$$

As you can see, the last row just shows all possible ways to represent m as the sum of two non-negative integers. Let's take a look at the coefficients:

$m=1$	$(a+b)^1$	$=$	a^1	$+$	b^1						
coefficients:			1		1						
$m=2$	$(a+b)^2$	$=$	a^2	$+$	$2a^1b^1$	$+$	b^2				
coefficients:			1		2		1				
$m=3$	$(a+b)^3$	$=$	a^3	$+$	$3a^2b^1$	$+$	$3a^1b^2$	$+$	b^3		
coefficients:			1		3		3		1		
$m=4$	$(a+b)^4$	$=$	a^4	$+$	$\boxed{?}a^3b^1$	$+$	$\boxed{?}a^2b^2$	$+$	$\boxed{?}a^1b^3$	$+$	b^4
coefficients:			1		4		?		4		1

We've made a guess for the last row, but the middle term remains unknown.

An actual computation reveals that it's 6. The table of coefficients becomes:

$m=1$	1	1			
$m=2$	1	2	1		
$m=3$	1	3	3	1	
$m=4$	1	4	6	4	1

An examination of two consecutive rows allows us to observe how things develop in this table:

$m=3$	1	$+$	3		3	$+$	3		3	$+$	1	
			$\|\|$				$\|\|$				$\|\|$	
$m=4$	1		4		4		6		6		4	

We can guess that every two adjacent coefficients in each row are added and placed in the next row.

The table is a *sequence of sequences*, and the terms in the next row are determined by the ones in the last. It is more convenient, however, to arrange the terms in an isosceles instead of a right triangle as above. This is how we build the 4th row from the 3rd:

known:	$m=3$			1				3				3				1	
			↙		↘		↙		↘		↙		↘		↙		↘
unknown:	$m=4$	1				1+3=4				3+3=6				3+1=4				1

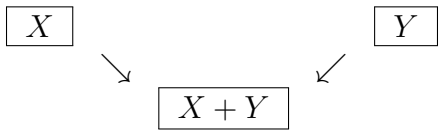
Each entry is the sum of the two entries above it. The result is called the *Pascal Triangle*:

$m=0$							1									
$m=1$						1		1								
$m=2$					1		2		1							
$m=3$				1		3		3		1						
$m=4$			1		4		6		4		1					
$m=5$		1		5		10		10		5		1				
...	.		.		.		.		.		.		.			

Exercise 1.7.7

Explain the entry at the top of the triangle.

The computation may continue indefinitely following this procedure:



We summarize the result below.

Theorem 1.7.8: Binomial Theorem

If two consecutive terms of the expansion of $(a+b)^m$ *are the following:*

$$X \cdot a^n b^{m-n} \quad \text{and} \quad Y \cdot a^{n-1} b^{m-n+1},$$

then the expansion of $(a+b)^{m+1}$ *contains the following term:*

$$(X+Y) \cdot a^n b^{m-n+1}.$$

Proof.

$$\begin{aligned}
&(Xa^n b^{m-n} + Ya^{n-1} b^{m-n+1})(a+b) \\
&= Xa^n b^{m-n} a + Ya^{n-1} b^{m-n+1} a + Xa^n b^{m-n} b + Ya^{n-1} b^{m-n+1} b \\
&= Xa^{n+1} b^{m-n} + Ya^n b^{m-n+1} + Xa^n b^{m-n+1} + Ya^{n-1} b^{m-n+2} \\
&= Xa^{n+1} b^{m-n} + (X+Y) a^n b^{m-n+1} + Ya^{n-1} b^{m-n+2}.
\end{aligned}$$

Exercise 1.7.9

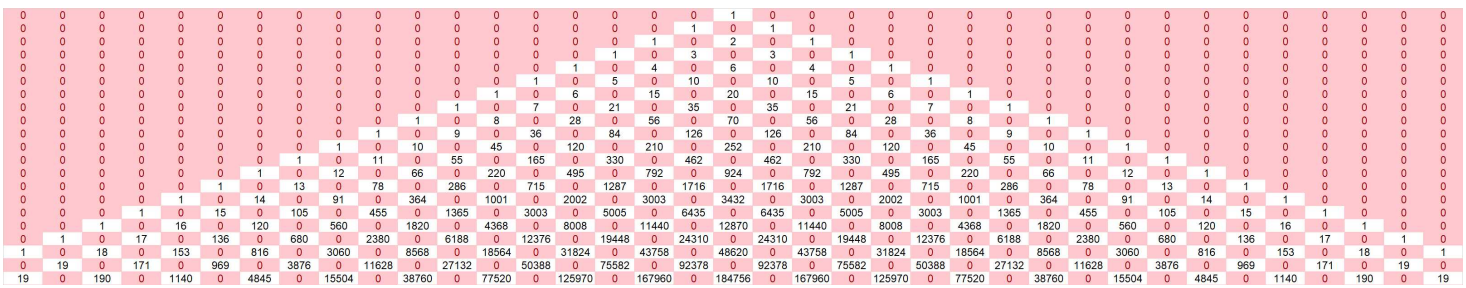
Provide the missing parts of the proof.

Example 1.7.10: Pascal Triangle, computed

This is how the Pascal Triangle is used. We take its 5th row and produce a new binomial expansion:

$$\begin{array}{lccccccccccc}
m=5 & & 1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
(a+b)^5 = & & 1\cdot a^5 & + & 5\cdot a^4 b & + & 10\cdot a^3 b^2 & + & 10\cdot a^2 b^3 & + & 5\cdot a b^4 & + & 1\cdot b^5
\end{array}$$

One can easily compute, recursively, a large part of the Pascal triangle with a spreadsheet:



We use the formula:

```
=R[-1]C[-1]+R[-1]C[1]
```

So, for each m, the coefficients of an expansion form a sequence of $m+1$ integers. They are simple in the very beginning and the very end:

n	0	1	...	$m-1$	m
	1	m	...	m	1

Let's examine the rest.

Definition 1.7.11: binomial coefficient

The nth term of this sequence is called the *binomial coefficient* and is denoted by

$$\boxed{\binom{m}{n}}$$

It reads "m choose n".

The first number gives the vertical location and the second horizontal location within the Pascal triangle.

A certain symmetry of the Pascal triangle is seen in the following formula.

Exercise 1.7.12

Show that

$$\binom{m}{n} = \binom{m}{m-n}.$$

Exercise 1.7.13

Show that

$$\binom{m}{2} = m\,.$$

In the new notation, the Pascal Triangle takes the form:

$$\begin{array}{ccccccccc} & & & & \binom{0}{0} & & & & \\ & & & \binom{1}{0} & & \binom{1}{1} & & & \\ & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & \\ & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & \\ . & & . & & . & & . & & . \end{array}$$

Therefore, the first number indicates the row, and the second number the position within the row.

In the new notation, the *Binomial Theorem* is restated as follows.

Corollary 1.7.14: Sum of Binomial Coefficients

For every positive integers m and $n \leq m$, we have:

$$\boxed{\binom{m}{n} + \binom{m}{n+1} = \binom{m+1}{n+1}}$$

That's a *recursive formula*! What is the *nth-term formula* for this sequence, or rather sequences? It can be expression in terms of the *factorial*. The following is an important result:

Theorem 1.7.15: Binomial Coefficient

For every positive integers m and $n \leq m$, we have:

$$\binom{m}{n} = \frac{m!}{n!(m-n)!}$$

Proof.

We confirm the recursive formula:

$$\begin{aligned}
\binom{m}{n} + \binom{m}{n+1} &= \frac{m!}{n!(m-n)!} + \frac{m!}{(n+1)!(m-n-1)!} \\
&= \frac{m!}{n!(m-n-1)!(m-n)} + \frac{m!}{n!(m-n-1)!(n+1)} \\
&= \frac{m!(n+1) + m!(m-n)}{n!(m-n-1)!(m-n)(n+1)} \\
&= \frac{m!\big((n+1)+(m-n)\big)}{(n+1)!(m-(n+1))!} \\
&= \frac{m!\big(m+1\big)}{(n+1)!(m-(n+1))!} \\
&= \binom{m+1}{n+1}.
\end{aligned}$$

Exercise 1.7.16

Why does this computation prove the formula?

Warning!

Even though a fraction, a binomial coefficient is an integer.

Example 1.7.17: binomial coefficients, computed

To test the formula, we substitute $m = 5$ and $n = 3$:

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{2 \cdot 3 \cdot 4 \cdot 5}{2 \cdot 3 \cdot 2} = 2 \cdot 5 = 10\,.$$

The result matches the one in the Pascal Triangle.

The binomial coefficients have another interpretation. Since the power is just a repeated multiplication,

$$(a+b)^m = \underbrace{(a+b) \cdot (a+b) \cdot \ldots \cdot (a+b)}_{m \text{ times}},$$

there will be one term in the binomial expansion for each choice of either a or b from each of the factors:

$$a^n b^{m-n} = \underbrace{a \cdot a \cdot \ldots \cdot a}_{n \text{ times}} \cdot \underbrace{b \cdot b \cdot \ldots \cdot b}_{m-n \text{ times}}.$$

We draw the following conclusion:

Theorem 1.7.18: Choose n From m

The binomial coefficient $\binom{m}{n}$ *is the number of ways we can choose* n *objects from* m *objects.*

Example 1.7.19: number of ways to choose a team

In how many ways can one form a team of 5 from 20 players? We compute:

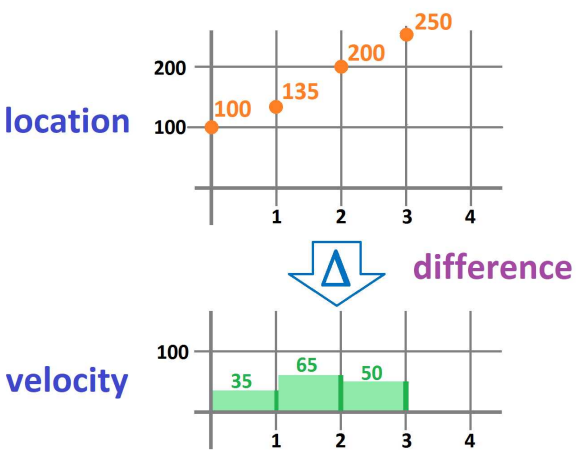
$$\begin{aligned}\binom{20}{5} &= \frac{20!}{5!15!}\\ &= \frac{15!\cdot 16\cdot 17\cdot 18\cdot 19\cdot 20}{5!15!}\\ &= \frac{16\cdot 17\cdot 18\cdot 19\cdot 20}{2\cdot 3\cdot 4\cdot 5}\\ &= \frac{16\cdot 17\cdot 18\cdot 19}{2\cdot 3}\\ &= 16\cdot 17\cdot 3\cdot 19\\ &= 15,504\,.\end{aligned}$$

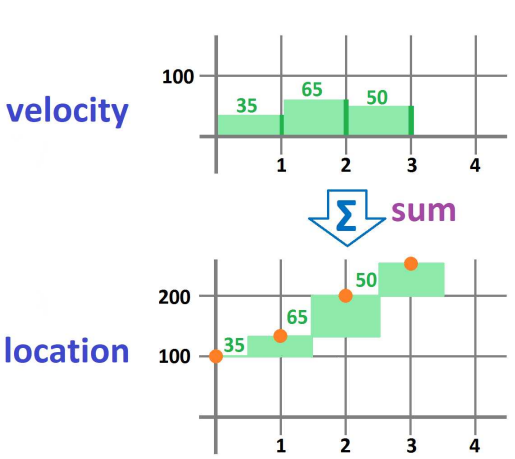
Exercise 1.7.20

How many different hands of 5 are there in a deck of 52 cards?

1.8. The sequence of differences: velocity

Numbers are subject to algebraic operations: addition, subtraction, multiplication, and division. Since the terms of sequences are numbers, a pair of sequences can be added (subtracted, etc.) to produce a new one. In addition, there are two operations that apply to a single sequence and produce a *new* sequence that tells us a lot about the *original* sequence. These operations are: subtracting the consecutive terms of the sequence and adding its terms repeatedly. We saw them in action in the first section:





This is the summary:

- If each term of a sequence represents a location, the pair-wise differences will give you the *velocities*, and
- If each term of a sequence represents the velocity, their sum up to that point will give you the *location*

(or displacement).

In this section we start on the path of development of an idea that culminates with the first fundamental concept of calculus, *the derivative* (Chapter 2DC-3).

The pairwise differences represent the *change* within the sequence, from each of its terms to the next.

Example 1.8.1: sequence given by list

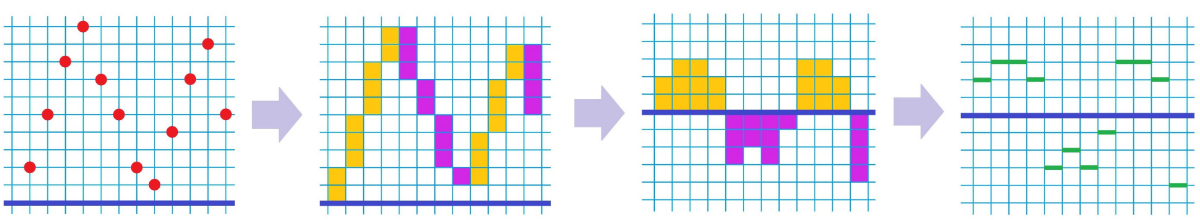
When a sequence is given by a list, we subtract the last term from the current one and put the result in the bottom row as follows:

a sequence:	2		4		7		1	...		
	↘	↙	↘	↙	↘	↙	↘	...		
its differences:		$4-2$		$7-4$		$1-7$		...		
		$\|\|$		$\|\|$		$\|\|$		...		
a new sequence:		2		3		-6		...		

We have a new list.

Example 1.8.2: sequence given by graph

In the simplest case, a sequence takes only integer values, then on the graph of the sequence, we just count the number of steps we make, up and down:



These increments then make a new sequence plotted on the right.

Definition 1.8.3: sequence of differences

For a sequence a_n, its *sequence of differences*, or simply the difference, is a new sequence, say d_n, defined for each n by the following:

$$d_n = a_{n+1} - a_n\,.$$

It is denoted by

$$\Delta a_n = a_{n+1} - a_n$$

Warning!

The symbol Δ applies to the *whole* sequence a_n, and Δa should be seen as the name of the new sequence; the notation for the difference is an abbreviation for $(\Delta a)_n$.

This is what the definition says:

Sequence of differences

a sequence:	a_1		a_2		a_3		a_4	...
		↘ ↙		↘ ↙		↘ ↙		...
its differences:		$a_2 - a_1$		$a_3 - a_2$		$a_4 - a_3$		...
		$\|\|$		$\|\|$		$\|\|$		...
a new sequence:		d_1		d_2		d_3		...
		$\|\|$		$\|\|$		$\|\|$		...
the notation:		Δa_1		Δa_2		Δa_3		...

Warning!

If the original sequence starts with $n = q$, then the new sequence starts with $n = q + 1$ but it could also be arranged to start with $n = q$ or any other index.

Now, what about sequences given by *formulas*? Let's consider a couple of specific sequences.

The first one is the *arithmetic progression* and it is very simple.

Theorem 1.8.4: Difference of Arithmetic Progression

The sequence of differences of an arithmetic progression with increment m is a constant sequence with the value equal to m.

Proof.

We simply compute from the definition:

$$d_n = \Delta(a_0 + mn) = a_{n+1} - a_n = (a_0 + b(m+1)) - (a_0 + mn) = m\,.$$

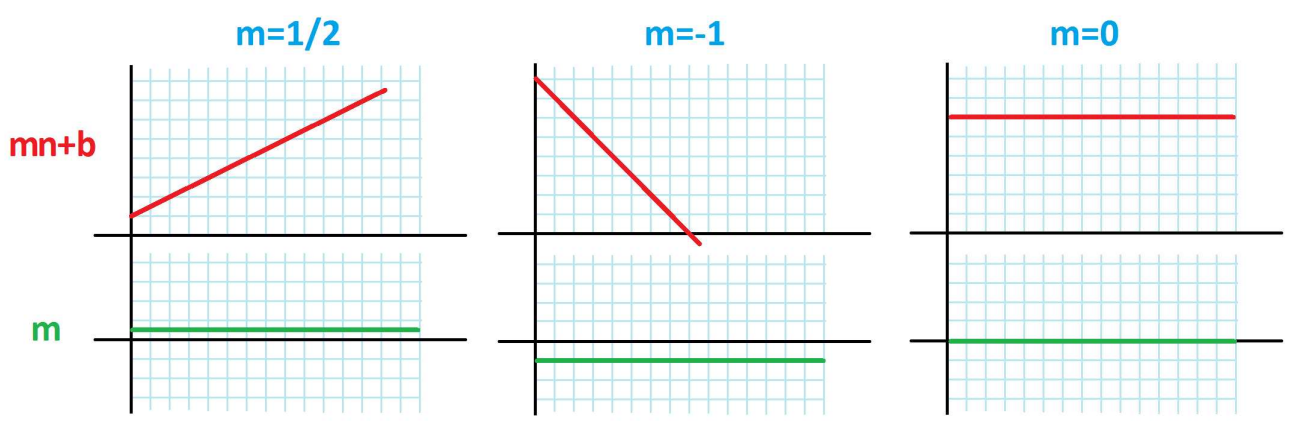
The theorem can be recast as an implication, an "if-then" statement:

▶ `IF` a_n is an arithmetic progression with increment m, `THEN` its sequence of differences is a constant sequence with the value equal to m.

We can also use our convenient abbreviation:

▶ a_n is an arithmetic progression with increment m $\implies$ its sequence of differences of a_n is a constant sequence with the value equal to m.

This is what the graphs of this pair of sequences may look like, zoomed out, for the following three choices of the increment m:



Let's take a look at a *geometric progression* $a_n = ar^n$ with $a > 0$ and $r > 0$.

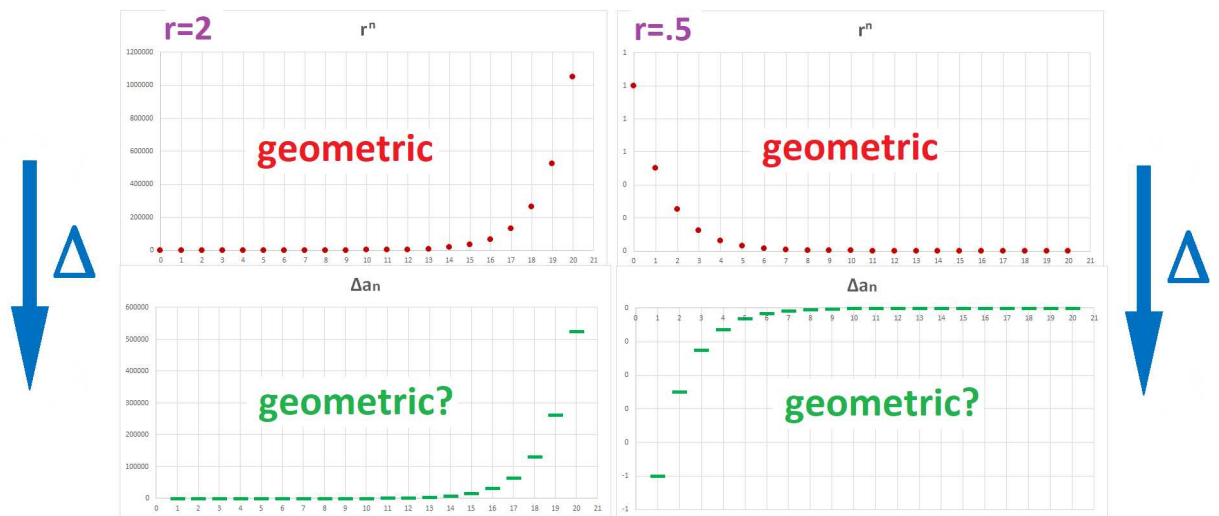
Example 1.8.5: geometric progression $a_n = 3^n$

This is a geometric progression with ratio $r = 3$. Let's compute the difference:

$$\begin{aligned}\Delta a_n &= a_{n+1} - a_n \\ &= 3^{n+1} - 3^n \\ &= 3^n \cdot 3 - 3^n \\ &= 3^n(3-1) \\ &= 3^n \cdot 2\,.\end{aligned}$$

It's a geometric progression with $r = 3$, again!

Is there a pattern? Let's plot the graph of a geometric progression $a_n = ar^n$ with $a > 0$ and $r > 0$. There are two cases, depending on the choice of ratio r (growth or decay):



What do the sequence of differences (second row) look like? We notice the following:

- It is positive and increasing, with speeding up when the ratio r is larger than 1.
- It is negative and increasing, with slowing down when $0 < r < 1$.

It also resembles the original sequence!

Theorem 1.8.6: Difference of Geometric Progression

The sequence of differences of a geometric progression is a geometric progression with the same ratio.

Proof.

If we have a geometric progression with ratio r and initial term a, its formula is $a_n = ar^n$. Therefore,

$$d_n = \Delta(ar^n) = a_{n+1} - a_n = ar^{n+1} - ar^n = a(r-1) \cdot r^n\,.$$

But that's the formula of a geometric progression with ratio r and initial term $a(r-1)$.

The theorem can be restated as an implication:

▶ `IF` a_n is a geometric progression with ratio r, `THEN` its sequence of differences is a geometric progression with ratio r.

Also:

▶ a_n is a geometric progression with ratio r $\implies$ its sequence of differences is a geometric progression with ratio r.

Example 1.8.7: alternating sequence

The sequence of differences of the alternating sequence $a_n = (-1)^n$ is computed below:

$$\Delta\left((-1)^n\right) = (-1)^{n+1} - (-1)^n = \begin{cases} (-1) - 1, & n \text{ is even} \\ 1 - (-1), & n \text{ is odd} \end{cases} = \begin{cases} -2, & n \text{ is even} \\ 2, & n \text{ is odd} \end{cases} = 2(-1)^{n+1}.$$

Exercise 1.8.8

What is the relation between the sequence above and its difference?

Example 1.8.9: differences are velocities

We can use computers to speed up these computations. For example, one may have been recording one's locations and now needs to find the velocities. Here is a spreadsheet formula for the sequence of differences (velocities):

```
=RC[-1]-R[-1]C[-1]
```

Whether the sequence comes from a formula or it's just a list of numbers, the formula applies equally:

=RC[-1]-R[-1]C[-1]

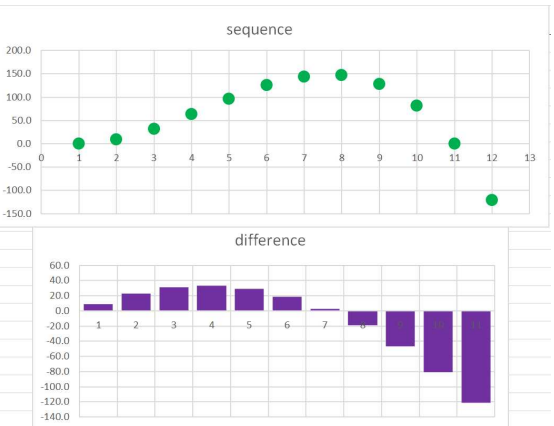
2	3	4
time	location	velocity
min	miles	miles/min
0	0.00	
1	0.10	=RC[-1]-R[-1]C[-1]
2	0.20	0.10

=RC[-1]-R[-1]C[-1]

2	3	4
time	location	velocity
min	miles	miles/min
0	0.00	
1	0.10	0.10
2	0.20	0.10

As a result, a curve has produced a new curve:

n	a_n	Δa_n
0	0.0	
1	9.0	9.0
2	32.0	23.0
3	63.0	31.0
4	96.0	33.0
5	125.0	29.0
6	144.0	19.0
7	147.0	3.0
8	128.0	-19.0
9	81.0	-47.0
10	0.0	-81.0
11	-121.0	-121.0



While the first graph tells us that we are moving forward and then backward, it is easier to derive better description from the second: speed up forward, then slow down, then speed up backward.

Exercise 1.8.10

Describe what has happened referring to, separately, the first graph and the second graph.

Exercise 1.8.11

Imagine, instead, that the first column of the spreadsheet above is where you have been recording the monthly balance of your bank account. What does the second column represent? Describe what has been happening with your finances referring to, separately, the first graph and the second graph.

This is the time for some *theory*.

Consider this obvious statement about motion:

- `IF` I am standing still, `THEN` my velocity is zero.

We can also say:

- ▶ `IF` my velocity is zero, `THEN` I am standing still.

We see the implications going both ways; the latter is the *converse* of the original statement (and vice versa!). Let's use symbols to restate these statements more compactly:

$$\begin{aligned} \text{I am standing still} &\implies \text{my velocity is zero.} \\ \text{I am standing still} &\impliedby \text{my velocity is zero.} \end{aligned}$$

The abbreviation of the combination of the two is an equivalence, an "if-and-only-if" statement:

$$\text{I am standing still} \iff \text{my velocity is zero.}$$

It's just another way of saying the same thing.

If we set the motion point of view aside, here is the general statement.

Theorem 1.8.12: Difference of Constant Sequence

A sequence is constant `IF AND ONLY IF` *its sequence of differences is zero.*

In other words, we have:

$$a_n \text{ is constant} \iff \Delta a_n = 0\,.$$

Proof.

Direct:

$$a_n = c \text{ for all } n \implies a_{n+1} - a_n = c - c = 0 \implies \Delta a_n = 0\,.$$

Converse:

$$a_{n+1} = a_n = c \text{ for all } n \impliedby a_{n+1} - a_n = 0 \impliedby \Delta a_n = 0\,.$$

Exercise 1.8.13

Prove that the difference of an arithmetic progression is constant and, conversely, that if the difference of a sequence is a constant sequence, then the sequence is an arithmetic progression.

Consider another obvious statement about motion:

- ▶ `IF` I am moving forward, `THEN` my velocity is positive.

And, conversely:

- ▶ `IF` my velocity is positive, `THEN` I am moving forward.

In other words, we have this pair of statements:

$$\begin{aligned} \text{Original:} \quad \text{I am moving forward} &\implies \text{my velocity is positive.} \\ \text{Converse:} \quad \text{I am moving forward} &\impliedby \text{my velocity is positive.} \end{aligned}$$

Exercise 1.8.14

What is the converse of the converse?

We combine these two in the following far reaching result.

Theorem 1.8.15: Monotonicity Theorem for Sequences

A sequence is increasing/decreasing `IF AND ONLY IF` *the sequence of differences is positive/negative or zero, respectively.*

In other words, we have:

$$\begin{array}{lll} a_n & \textit{is increasing} & \iff \Delta a_n \geq 0\,. \\ a_n & \textit{is decreasing} & \iff \Delta a_n \leq 0\,. \\ a_n & \textit{is constant} & \iff \Delta a_n = 0\,. \end{array}$$

Proof.

$$a_{n+1} \geq a_n \text{ for all } n \iff a_{n+1} - a_n \geq 0 \iff \Delta a_n \geq 0\,.$$

It's just another way of saying the same thing.

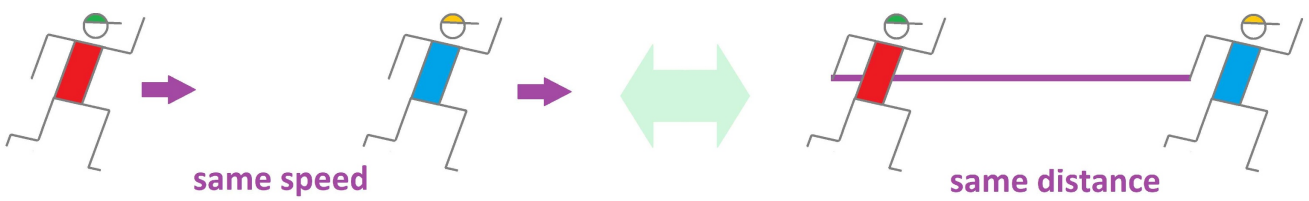
Suppose now that there are *two* runners; then we have a less obvious fact about motion:

▶ `IF` the distance between two runners isn't changing, `THEN` they are running with the same velocity.

And vice versa:

▶ `IF` two runners are running with the same velocity, `THEN` the distance between them isn't changing.

It's as if they are holding the two ends of a pole:



The conclusion holds even if they speed up and slow down all the time. In other words, we have:

▶ The distance between two runners isn't changing `IF AND ONLY IF` they are running with the same velocity.

Once again, for sequences a_n and b_n representing their respective positions at time n, we can restate this idea mathematically in order to confirm that our theory makes sense.

Corollary 1.8.16: Difference under Subtraction

Two sequences differ by a constant `IF AND ONLY IF` *if their sequences of differences are equal.*

In other words, we have:

$$a_n - b_n \textit{ is constant} \iff \Delta a_n = \Delta b_n\,.$$

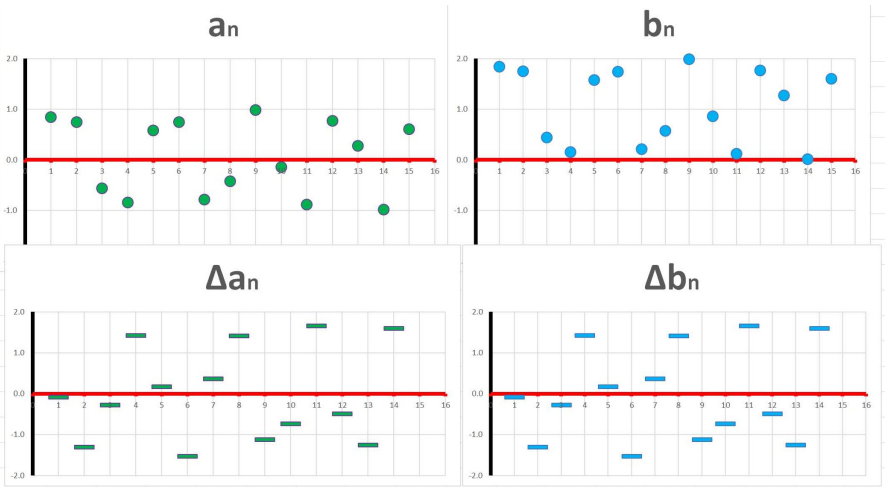
Proof.

The corollary follows from the *Difference of Constant Sequence Theorem* above.

Example 1.8.17: shift of sequence

We shift the sequence a_n below by 1 unit up to produce a new sequence b_n (top):

n	a_n	b_n	Δa_n	Δb_n
1	0.84	1.84		
2	0.75	1.75	-0.09	-0.09
3	-0.56	0.44	-1.31	-1.31
4	-0.84	0.16	-0.28	-0.28
5	0.58	1.58	1.42	1.42
6	0.74	1.74	0.17	0.17
7	-0.79	0.21	-1.53	-1.53
8	-0.43	0.57	0.36	0.36
9	0.99	1.99	1.41	1.41
10	-0.14	0.86	-1.13	-1.13
11	-0.88	0.12	-0.74	-0.74
12	0.77	1.77	1.65	1.65
13	0.27	1.27	-0.49	-0.49
14	-0.99	0.01	-1.26	-1.26
15	0.60	1.60	1.59	1.59



Because the ups and downs remain the same, the sequences of differences of these two sequences are identical (bottom).

Exercise 1.8.18

What if the two runners holding the pole also start to move their hands back and forth?

We can use the latter theorem to watch after the distance between the two runners. A matching statement about motion is the following:

▶ `IF` the distance from one of the two runners to the other is increasing, `THEN` the former's velocity is higher.

Conversely:

▶ `IF` the velocity of one runner is higher than the other, `THEN` the distance between them is increasing.

Exercise 1.8.19

Combine the two statements into one.

We can restate this mathematically.

Corollary 1.8.20: Monotonicity and Subtraction

The difference of two sequences is increasing `IF AND ONLY IF` *the former's difference is bigger than the latter's.*

In other words, we have:

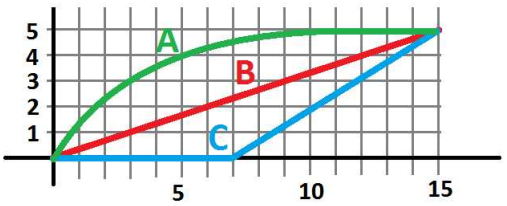
$$\begin{aligned} a_n - b_n \quad &\text{is increasing} \quad \iff \quad \Delta a_n \geq \Delta b_n\,, \\ a_n - b_n \quad &\text{is decreasing} \quad \iff \quad \Delta a_n \leq \Delta b_n\,. \end{aligned}$$

Proof.

The corollary follows from the *Monotonicity Theorem for Sequences* above.

Example 1.8.21: three runners

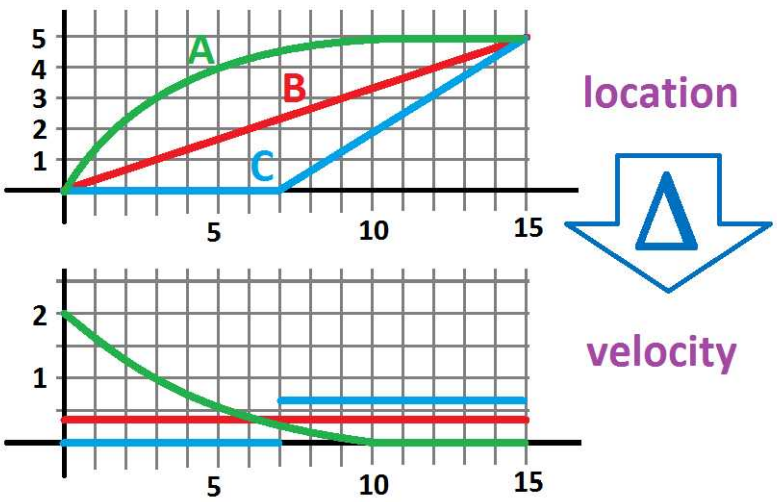
The graph below shows the positions of three runners in terms of time, n. Describe what has happened:



They are all at the starting line together, and at the end, they are all at the finish line. Furthermore, A reaches the finish line first, followed by B, and then C (who also starts late). This is *how* each did it:

- A starts fast, then slows down, and almost stops close to the finish line.
- B maintains the same speed.
- C starts late and then runs fast at the same speed.

We can see that A is running faster because the distance from B is increasing. It becomes slower later, which is visible from the decreasing distance. We can discover this and the rest of the facts by examining the graphs of the *differences* of the sequences:



Exercise 1.8.22

Suppose a sequence is given by the graph for velocity above. Sketch the graph of the difference of this sequence. What is its meaning?

Exercise 1.8.23

Plot the location and the velocity for the following trip: "I drove fast, then gradually slowed down, stopped for a very short moment, gradually accelerated, maintained speed, hit a wall." Make up your own story and repeat the task.

Exercise 1.8.24

Draw a curve on a piece of paper, imagine that it represents your locations, and then sketch what your velocity would look like. Repeat.

Exercise 1.8.25

Imagine that the first graph represents, instead of locations, the balances of three bank accounts. Describe what has been happening.

Example 1.8.26: difference quotient of sequence

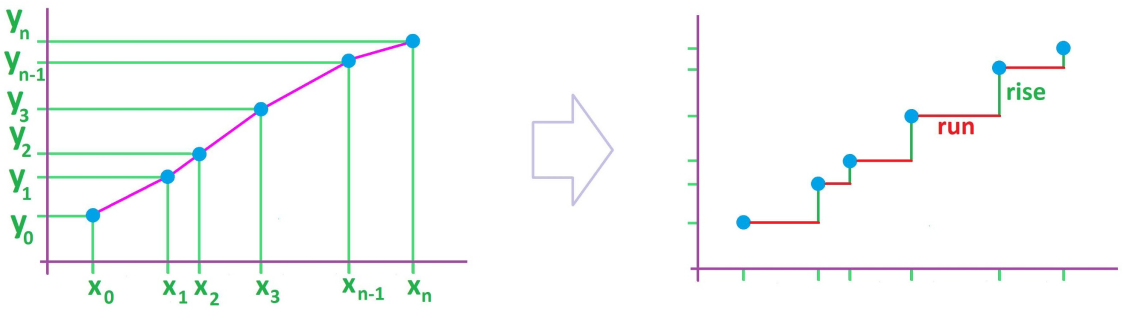
How do we treat motion when the time increment isn't 1? What is the *velocity* then?

First, we define the time and the location by two separate sequences, say, x_n and y_n. Then the

velocity is the increment of the latter over the increment of the former. We notice that those two are the differences of the two sequences. The *difference quotient* of two sequences of x_n and y_n is defined to be the sequence that is the difference of y_n divided by the difference of x_n:

$$\frac{\Delta y_n}{\Delta x_n} = \frac{y_{n+1} - y_n}{x_{n+1} - x_n},$$

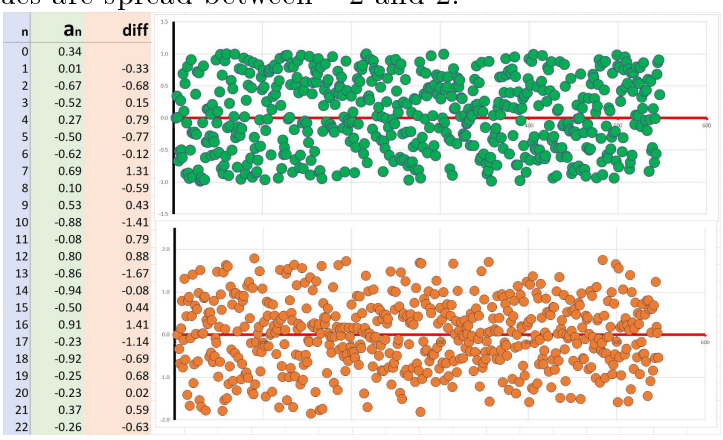
provided the denominator is not zero. It is the relative change – the *rate* of change – of the two sequences (for each consecutive pair of points, it is the slope):



The difference quotient of the sequence of the location with respect to the sequence of time is the velocity. (We will continue this study in Chapter 2.)

Example 1.8.27: difference of random sequence

If we take a sequence of random numbers, with its values spread between -1 and 1, its difference is also random but the values are spread between -2 and 2:



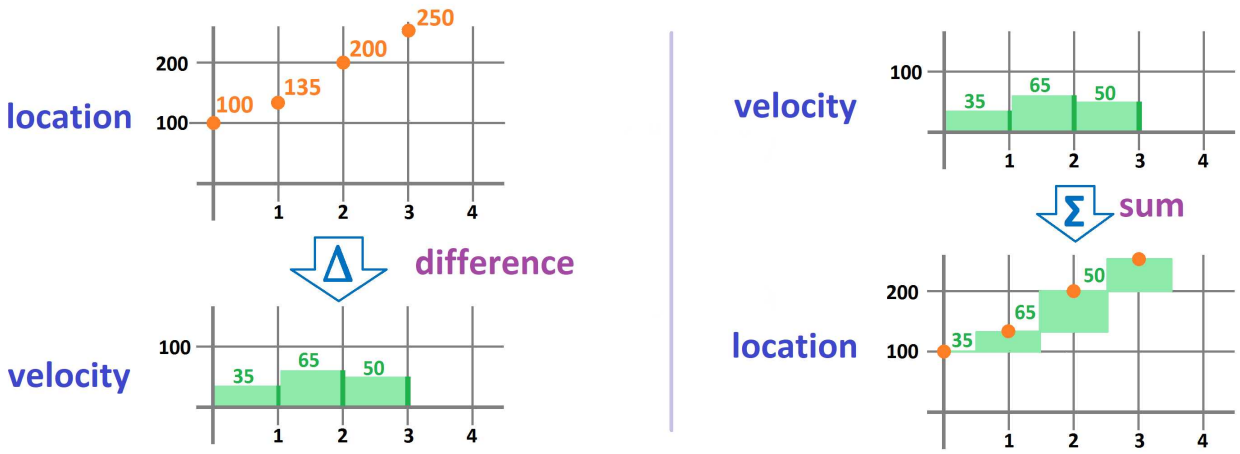
n	a_n	diff
0	0.34	
1	0.01	-0.33
2	-0.67	-0.68
3	-0.52	0.15
4	0.27	0.79
5	-0.50	-0.77
6	-0.62	-0.12
7	0.69	1.31
8	0.10	-0.59
9	0.53	0.43
10	-0.88	-1.41
11	-0.08	0.79
12	0.80	0.88
13	-0.86	-1.67
14	-0.94	-0.08
15	-0.50	0.44
16	0.91	1.41
17	-0.23	-1.14
18	-0.92	-0.69
19	-0.25	0.68
20	-0.23	0.02
21	0.37	0.59
22	-0.26	-0.63

Exercise 1.8.28

Point out and explain the difference between the two graphs.

1.9. The sequence of the sums: displacement

In the first section, we saw how the sequences of *locations and velocities* interact. We took a closer look at the transition from the former to the latter and now in reverse:



In this section, we start on the path of development of an idea that culminates with the second fundamental concept of calculus, *the integral* (Chapter 3IC-1).

The sum represents the totality of the "beginning" of a sequence, found by adding each of its terms to the next, up to that point.

Example 1.9.1: sequences given by lists

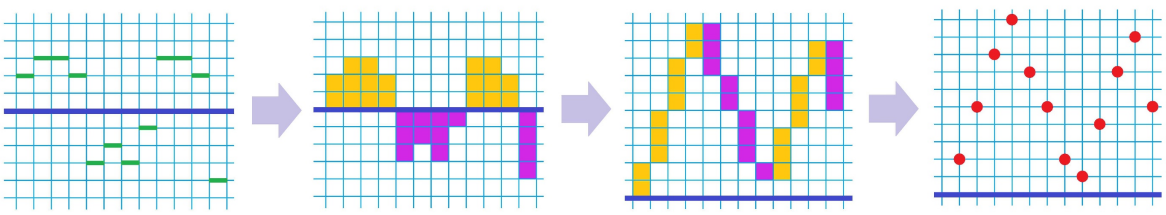
We just add the current term to what we have accumulated so far:

sequence:	2		4			7			1			-1			...			
	↓		↓			↓			↓			↓			...			
sums:	2																	
	2	+	4	=	6													
					6	+	7	=	13									
									13	+	1	=	14					
													14	+	(-1)	=	13	
	↓				↓				↓				↓				↓	...
new sequence:	2				6				13				14				13	...

We have a new list!

Example 1.9.2: sequences given by graphs

We treat the graph of a sequence as if made of bars and then just stack up these bars on top of each other one by one:



These stacked bars – or rather the process of stacking – make a new sequence.

Unlike the difference, the sum must be defined (and computed) in a *recursive* manner.

Definition 1.9.3: sequence of sums

For a sequence a_n, its *sequence of sums*, or simply the sum, is a new sequence s_n defined and denoted for each $n \geq m$ within the domain of a_n by the following (recursive) formula:

$$s_m = 0, \quad s_{n+1} = s_n + a_{n+1}$$

In other words, we have:

$$s_n = a_m + a_{m+1} + ... + a_n$$

Example 1.9.4: alternating sequence

Let's do some algebra. Here a_n is the original sequence and s_n is the new one:

n	a_n	s_n	$=$	s_n
1	1	1	$=$	1
2	-1	$1-1$	$=$	0
3	1	$1-1+1$	$=$	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	$(-1)^n$	$1-1+1-...+(-1)^n$	$=$	1 or 0

The resulting sequence is also "alternating"!

A commonly used notation is the following:

Sigma notation for summation

$$s_n = a_m + a_{m+1} + ... + a_n = \sum_{k=m}^{n} a_k$$

Let's take a closer look at the new notation. The first choice of how to represent the sum of a segment – from m to n – of a sequence a_n is this:

$$\underbrace{a_m}_{\text{step 1}} \quad \underbrace{+a_{m+1}}_{\text{step 2}} +... \quad \underbrace{+a_k}_{\text{step } k} +... \quad \underbrace{+a_n}_{\text{step } n-m} .$$

This notation reflects the recursive nature of the process but it can also be repetitive and cumbersome. The second choice is more compact:

$$\sum_{k=m}^{n} a_k \, .$$

Here the Greek letter Σ stands for the letter S meaning "sum".

Sigma notation

$$\sum_{k=0}^{3} \left(k^2+k\right) = 20 \quad \longrightarrow \quad \begin{array}{c} \text{beginning} \\ \downarrow \\ 3 \\ \sum \\ k=0 \end{array} \begin{array}{c} \text{and end values for } k \\ \\ \left(\quad k^2+k \quad \right) \\ \uparrow \\ \text{a specific sequence} \end{array} \quad = \quad \begin{array}{c} \\ \\ 20 \\ \uparrow \\ \text{a specific number} \end{array}$$

Warning!

It would make sense to have "$k = 0$" above the sigma:

$$\sum_{k=0}^{k=3} \left(k^2+k\right) .$$

Example 1.9.5: expanding from sigma notation

The computation above is expanded here:

$$\sum_{k=0}^{3}\left(k^2+k\right) = \begin{array}{c|ll|l} k & k^2+k & & \\ \hline 0 & 0^2+0 & =0 & + \\ 1 & 1^2+1 & =2 & + \\ 2 & 2^2+2 & =6 & + \\ 3 & 3^2+3 & =12 & \\ \hline & & =20 & \end{array}$$

Exercise 1.9.6: contracting to sigma notation

How will the sum change if we replace $k=0$ with $k=1$, or $k=-1$? What if we replace 3 at the top with 4?

Example 1.9.7: contracting summation

This is how we *contract* the summation:

$$1^2+2^2+3^2+...+17^2=\sum_{k=1}^{n}k^2\,.$$

This is only possible if we find the nth-term formula for the sequence; in this case, $a_k=k^2$. And this is how we *expand* back from this compact notation, by plugging the values of $k=1,2,...,17$ into the formula:

$$\sum_{k=1}^{17}k^2=\underbrace{1^2}_{k=1}+\underbrace{2^2}_{k=2}+\underbrace{3^2}_{k=3}+...+\underbrace{17^2}_{k=17}\,.$$

Similarly, we have:

$$1+\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+...+\frac{1}{2^{10}}=\sum_{k=0}^{10}\frac{1}{2^k}\,.$$

Exercise 1.9.8

Confirm that we can start at any other initial index if we just modify the formula:

$$1+\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+...+\frac{1}{2^{10}}=\sum_{k=?}^{?}\frac{1}{2^{k-1}}=\sum_{k=?}^{?}\frac{1}{2^{k-2}}=...$$

Exercise 1.9.9

Contract this summation:

$$1+\frac{1}{3}+\frac{1}{9}+\frac{1}{27}=\,?$$

Exercise 1.9.10

Expand this summation:

$$\sum_{k=0}^{4}(k/2)=\,?$$

Exercise 1.9.11

Rewrite using the sigma notation:

1. $1+3+5+7+9+11+13+15$
2. $.9+.99+.999+.9999$
3. $1/2-1/4+1/8-1/16$
4. $1+1/2+1/3+1/4+...+1/n$
5. $1+1/2+1/4+1/8$
6. $2+3+5+7+11+13+17$
7. $1-4+9-16+25$

Example 1.9.12: binomials

In this notation, the *Binomial Theorem* reads:

$$(a+b)^m = \sum_{n=0}^{m} \binom{m}{n} a^{m-n} b^n \,.$$

For example:

$$(a+b)^4 = \sum_{n=0}^{4} \binom{4}{n} a^{4-n} b^n \,.$$

Then we have:

$$\begin{array}{rrcrcrcrcr}
= & \binom{4}{0}a^{4-0}b^0 & + & \binom{4}{1}a^{4-1}b^1 & + & \binom{4}{2}a^{4-2}b^2 & + & \binom{4}{3}a^{4-3}b^3 & + & \binom{4}{4}a^{4-4}b^4 \\
= & 1\cdot a^4b^0 & + & 4\cdot a^3b^1 & + & 6\cdot a^2b^2 & + & 4\cdot a^1b^3 & + & 1\cdot a^0b^4 \\
= & a^4 & + & 4a^3b & + & 6a^2b^2 & + & 4ab^3 & + & b^4 \,.
\end{array}$$

The notation applies to all sequences, both finite and infinite. For infinite sequences, recognized by "..." at the end, the sum sequence is also called "partial sums" as well as "series" (to be discussed in Chapter 3IC-5).

This is the definition of the sequence of sums written with the sigma notation:

Sequence of sums

$$\begin{array}{rccccccccccc}
\text{a sequence:} & a_1 & & a_2 & & & a_3 & & & a_4 & & \dots \\
& \downarrow & & \downarrow & & & \downarrow & & & \downarrow & & \dots \\
\text{its sums:} & a_1 & & & & & & & & & & \\
& a_1 & + & a_2 & = & s_2 & & & & & & \\
& & & & & s_2 & + & a_3 & = & s_3 & & \\
& & & & & & & & & s_3 & + & a_4 \\
& & & & & & & & & & & \\
& \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \dots \\
\text{the sequence of sums:} & s_1 & & s_2 & & s_3 & & s_4 & & & & \dots \\
& \| & & \| & & \| & & \| & & & & \dots \\
\text{the sigma notation:} & \sum_{k=1}^{1} a_k & & \sum_{k=1}^{2} a_k & & \sum_{k=1}^{3} a_k & & \sum_{k=1}^{4} a_k & & & & \dots
\end{array}$$

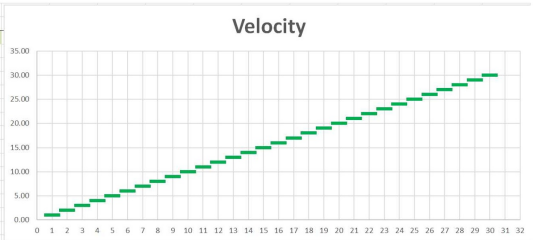
Below is the simplest result about the sums. It is still considerably more challenging than most results about the differences that we saw in the last section.

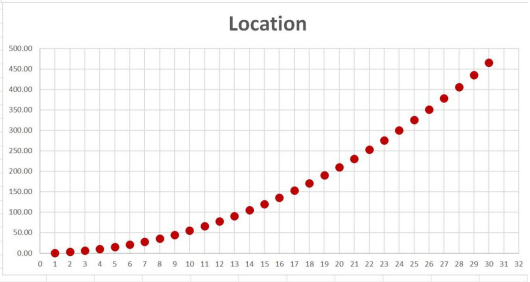
Theorem 1.9.13: Sum of Arithmetic Progression

The sum of an arithmetic progression with increment m and a 0 initial term is a (quadratic) sequence given by the following:

$$\sum_{k=1}^{n}(mk) = \frac{n(n+1)}{2}m\,.$$

time min	velocity miles/min	location miles
0		0.00
1	1.00	1.00
2	2.00	3.00
3	3.00	6.00
4	4.00	10.00
5	5.00	15.00
6	6.00	21.00
7	7.00	28.00
8	8.00	36.00
9	9.00	45.00
10	10.00	55.00
11	11.00	66.00
12	12.00	78.00
13	13.00	91.00
14	14.00	105.00
15	15.00	120.00
16	16.00	136.00
17	17.00	153.00
18	18.00	171.00
19	19.00	190.00
20	20.00	210.00
21	21.00	231.00
22	22.00	253.00
23	23.00	276.00
24	24.00	300.00
25	25.00	325.00
26	26.00	351.00
27	27.00	378.00
28	28.00	406.00
29	29.00	435.00
30	30.00	465.00





Exercise 1.9.14

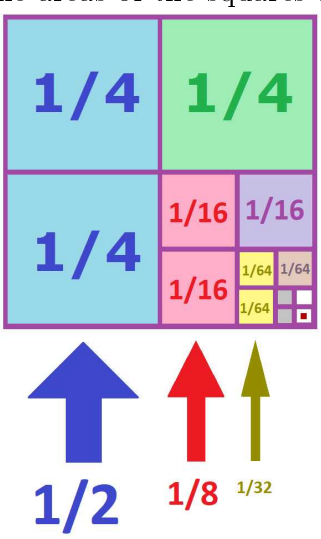
Prove the theorem. Hint: Use the example about round robins presented earlier in this chapter.

Example 1.9.15: from sequences to series

What does the sum

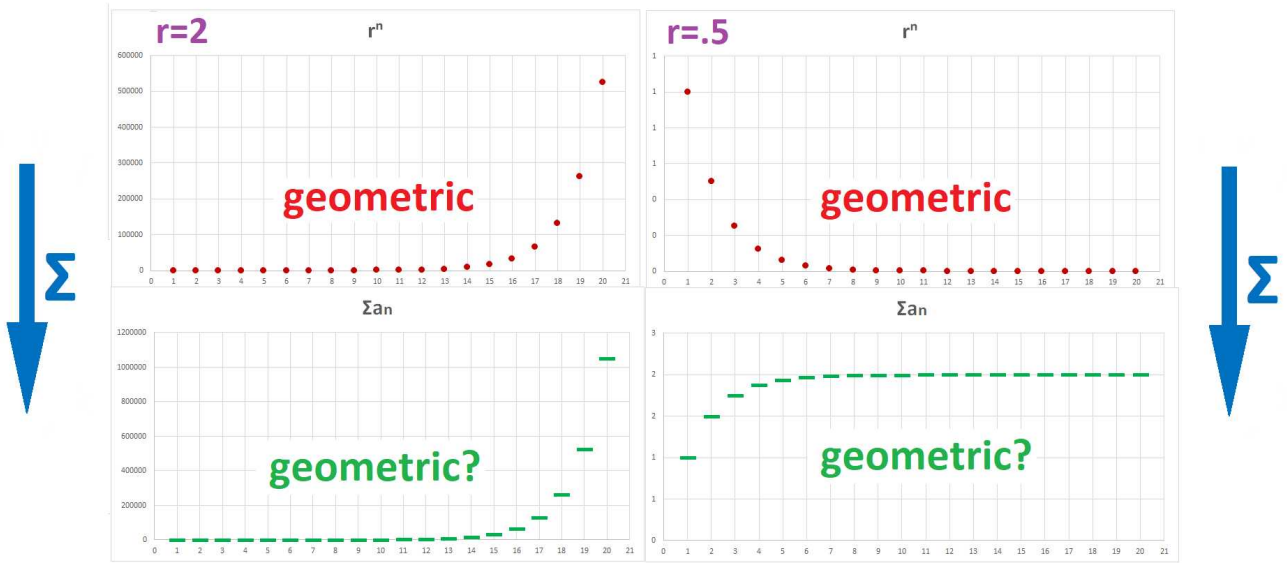
$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots$$

compute? We can see these terms as the areas of the squares below:



On the one hand, adding the areas of these squares will never go over 1 (the area of the big square), and, on the other, these squares seem to exhaust this square entirely. So, even the *infinite* sum sometimes makes sense (to be considered in Chapter 3IC-5). Of course, this is a geometric progression.

Let's take a look at a *geometric progression* with $a_n = ar^n$ with $a > 0$ and $r > 0$. There are two cases, depending on the choice of ratio r (growth or decay):



What about the sequence of sums (second row)? We notice the following:

- It is increasing with speeding up when its ratio r is larger than 1.
- It is increasing with slowing down when $0 < r < 1$.

It also resembles the original sequence!

Theorem 1.9.16: Sum of Geometric Progression

The sequence of sums of a geometric progression with ratio $r \neq 1$ is a geometric progression with the same ratio and a constant sequence.

In other words, we have:

$$\sum_{k=1}^{n} ar^k = Ar^n + C\,,$$

for some real numbers A and C.

Proof.

Below, we use a *clever trick* to get rid of "...". We write the nth sum s_n,

$$s_n \;= ar^0 \;+ar^1 \;+ar^2 \;+\ldots\; +ar^{n-1} \;+ar^n$$

and then multiply it by r:

$$\begin{aligned} rs_n \;&= r\big(ar^0 \;+ar^1 \;+ar^2 \;+\ldots\; +ar^{n-1} \;+ar^n\big) \\ &= \;\; ar^1 \;+\; ar^2 \;+\; ar^3 \;+\ldots\; +\; ar^n \;+\; ar^{n+1} \end{aligned}$$

Now we subtract these two:

$$\begin{array}{lllllll} s_n & = ar^0 & +ar^1 & +ar^2 & +\ldots & +ar^{n-1} & +ar^n \\ rs_n & = \;\; ar^1 & +\;\; ar^2 & +\;\; ar^3 & +\ldots & +\;\; ar^n & +\;\; ar^{n+1} \\ \hline s_n - rs_n & = ar^0 - ar^1 & +ar^1 - ar^2 & +ar^2 - ar^3 & +\ldots & +ar^{n-1} - ar^n & +ar^n - ar^{n+1} \\ & = ar^0 & & & & & -ar^{n+1} \end{array}$$

We cancel the terms that appear twice in the last row and "..." is gone! Therefore,

$$s_n(1-r) = a - ar^{n+1}\,.$$

Thus, we have an explicit formula for the nth term of the sum:

$$s_n = \frac{a}{1-r}(1-r^{n+1}) = -\frac{a}{1-r}\cdot r^{n+1} + \frac{a}{1-r}\,.$$

The former term is the geometric part, and the latter is the constant:

$$A = -\frac{ar}{1-r},\ C = \frac{a}{1-r}\,.$$

Exercise 1.9.17

Find the explicit formula for the sequence of sums of the alternating sequence $a_n = (-1)^n$.

Exercise 1.9.18

Use the trick to prove the theorem about arithmetic progressions.

Warning!

Our ability to produce an explicit formula for the nth terms of the sequence of the sum is an exception, not a rule.

Example 1.9.19: sums are displacements

We can use computers to speed up these computations. For example, one may have been recording one's velocities and now looking for the location. This is a formula for a spreadsheet (the locations):

```
=R[-1]C+RC[-1]
```

Whether the sequence comes from a formula or it's just a list of numbers, the formula applies:

=R[-1]C+RC[-1]

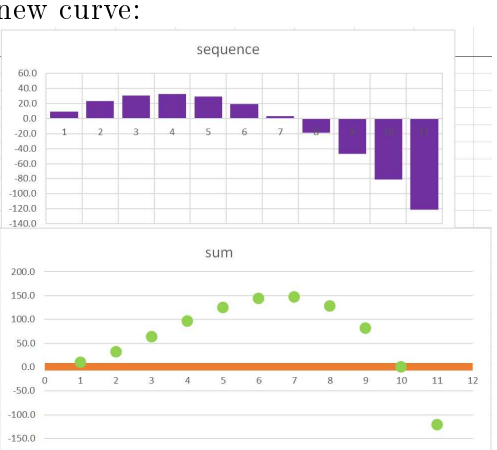
2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	=R[-1]C+RC[-1]
2	0.20	0.30

=R[-1]C+RC[-1]

2	3	4
time	velocity	location
min	miles/min	miles
0		0.00
1	0.10	0.10
2	0.20	0.30

As a result, a curve has produced a new curve:

a_n	Σa_n
	0.0
9.0	9.0
23.0	32.0
31.0	63.0
33.0	96.0
29.0	125.0
19.0	144.0
3.0	147.0
-19.0	128.0
-47.0	81.0
-81.0	0.0
-121.0	-121.0



Exercise 1.9.20

Describe what has happened referring to, separately, the first graph and the second graph.

Exercise 1.9.21

Imagine that the first column of the spreadsheet is where you have been recording your monthly deposit/withdrawals at your bank account. What does the second column represent? Describe what

has been happening referring to, separately, the first graph and the second graph.

This is the time for some *theory*.

Recall from the last section this pair of obvious statements about motion:

- ▶ I am standing still `IF AND ONLY IF` my velocity is zero.

If the velocity is represented by a sequence, its sum is the location. We can then restate the above mathematically.

Theorem 1.9.22: Constant Sequence as Sum

The sequence of sums of a sequence is constant `IF AND ONLY IF` *the sequence has only zero values starting from some index N.*

In other words, we have:

$$\sum_{k=m}^{n} a_k \text{ is constant} \iff a_n = 0 \text{ for all } n \geq N\,.$$

Proof.

$$\sum_{k=m}^{n} a_k = c \text{ for all } n \iff a_{n+1} = \sum_{k=m}^{n+1} a_k - \sum_{k=m}^{n} a_k = c - c = 0 \iff a_{n+1} = 0\,.$$

Here is another equivalence statements about motion:

- ▶ I am moving forward `IF AND ONLY IF` my velocity is positive.

We can restate this mathematically using the sums.

Theorem 1.9.23: Monotonicity of Sum

The sequence of sums of a sequence is increasing `IF AND ONLY IF` *the terms of the sequence are non-negative.*

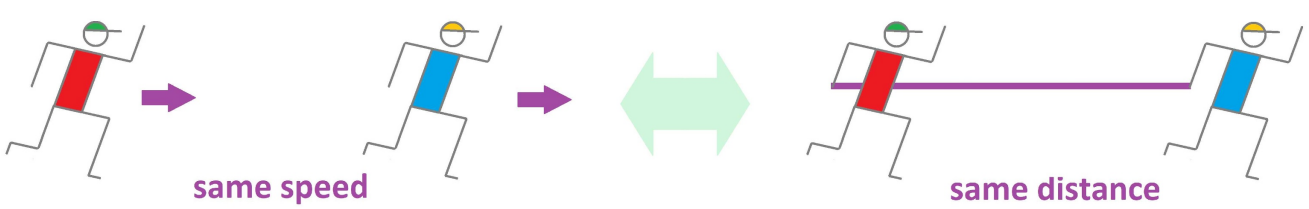
In other words, we have:

$$\sum_{k=m}^{n} a_k \text{ is increasing} \iff a_n \geq 0\,.$$
$$\sum_{k=m}^{n} a_k \text{ is decreasing} \iff a_n \leq 0\,.$$

Proof.

$$\sum_{k=m}^{n+1} a_k \geq \sum_{k=m}^{n} a_k \text{ for all } n \iff a_{n+1} = \sum_{k=m}^{n+1} a_k - \sum_{k=m}^{n} a_k \geq 0\,.$$

Now suppose just like in the last section, that there are *two* runners:



Then, we have:

- the distance between two runners isn't changing `IF AND ONLY IF` they are running with the same velocity.

We restate this mathematically.

Corollary 1.9.24: Subtracting Sums of Sequences

The sequences of sums of two sequences differ by a constant `IF AND ONLY IF` *the sequences are equal starting with some term N.*

In other words, we have:

$$\sum_{k=m}^{n} a_k - \sum_{k=m}^{n} b_k \text{ is constant} \iff a_n = b_n \text{ for all } n \geq N\,.$$

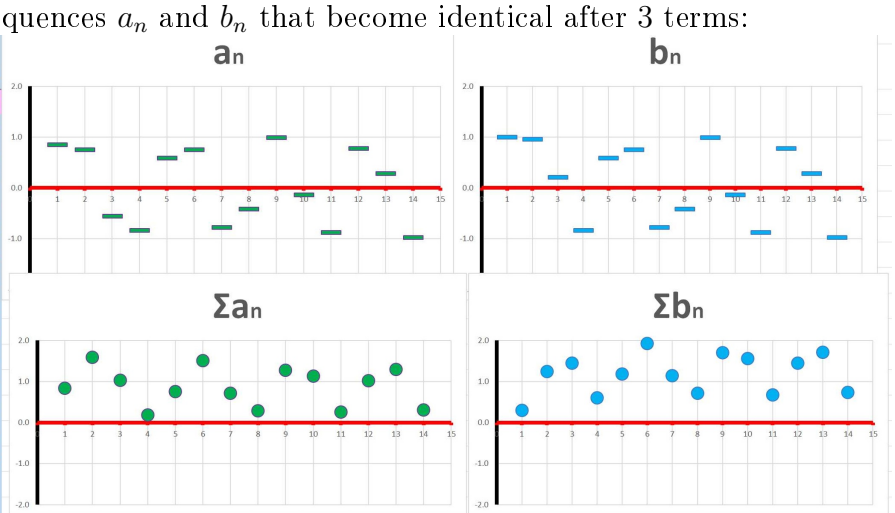
Proof.

The corollary follows from the *Constant Sequence as Sum* above.

Example 1.9.25: shift of sequence

We have below two different sequences a_n and b_n that become identical after 3 terms:

n	a_n	b_n	Σa_n	Σb_n
0			0.00	0.00
1	0.84	0.99	0.84	0.30
2	0.75	0.95	1.59	1.25
3	-0.56	0.20	1.03	1.45
4	-0.84	-0.84	0.18	0.61
5	0.58	0.58	0.76	1.18
6	0.74	0.74	1.50	1.93
7	-0.79	-0.79	0.72	1.14
8	-0.43	-0.43	0.29	0.71
9	0.99	0.99	1.28	1.70
10	-0.14	-0.14	1.14	1.56
11	-0.88	-0.88	0.25	0.68
12	0.77	0.77	1.02	1.44
13	0.27	0.27	1.30	1.72
14	-0.99	-0.99	0.31	0.73



The result is that the sum of the latter sequence is just a vertical shift of the sum of the former:

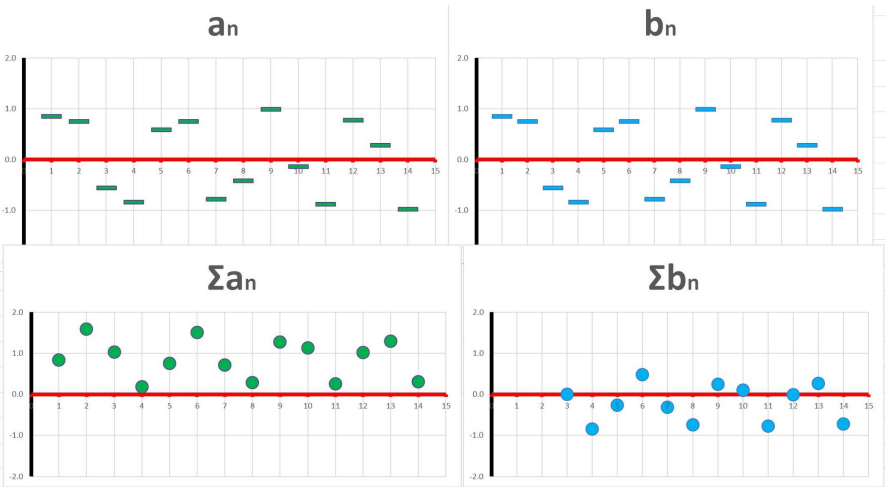


To state this algebraically, we have for each n:

$$\sum_{k=m}^{n} a_k = \sum_{k=m}^{n} b_k + C\,,$$

where C is some number. The outcome is the same when the two sequences a_n and b_n are identical but the computation of their sequences of sums starts at different points:

n	a_n	b_n	Σa_n	Σb_n
0			0.00	
1	0.84	0.84	0.84	
2	0.75	0.75	1.59	
3	-0.56	-0.56	1.03	0.00
4	-0.84	-0.84	0.18	-0.84
5	0.58	0.58	0.76	-0.27
6	0.74	0.74	1.50	0.48
7	-0.79	-0.79	0.72	-0.31
8	-0.43	-0.43	0.29	-0.74
9	0.99	0.99	1.28	0.25
10	-0.14	-0.14	1.14	0.11
11	-0.88	-0.88	0.25	-0.77
12	0.77	0.77	1.02	-0.01
13	0.27	0.27	1.30	0.27
14	-0.99	-0.99	0.31	-0.72



Exercise 1.9.26

What is the meaning of the number C?

We can use the theorems to watch for the distance between the two runners:

- ▶ the distance from one of the two runners to the other is increasing `IF AND ONLY IF` the former's velocity is higher.

We can restate this mathematically using the sums.

Corollary 1.9.27: Subtracting Sums: Monotonicity

The difference of the sequences of sums of two sequences is increasing `IF AND ONLY IF` *the corresponding terms of the former are larger than or equal to those of the latter.*

In other words, we have:

$$\sum_{k=m}^{n} a_k - \sum_{k=m}^{n} b_k \quad \textit{is increasing} \quad \Longleftrightarrow \quad a_n \geq b_n\,.$$

$$\sum_{k=m}^{n} a_k - \sum_{k=m}^{n} b_k \quad \textit{is decreasing} \quad \Longleftrightarrow \quad a_n \leq b_n\,.$$

Proof.

The corollary follows from *Monotonicity of Sum* above.

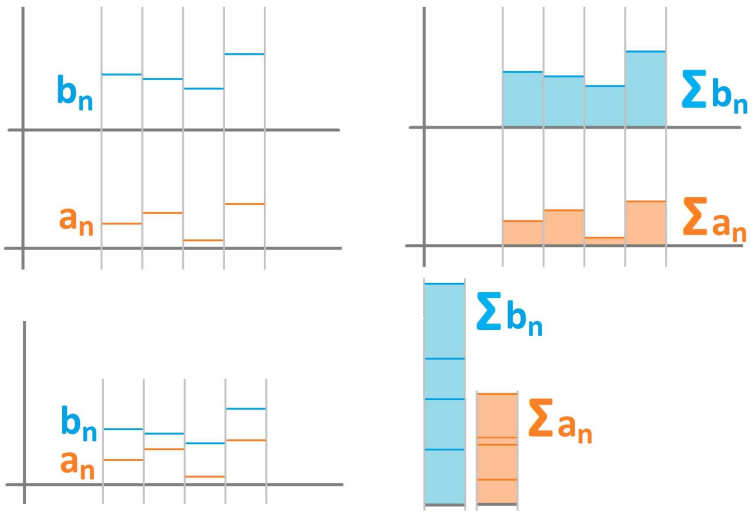
Here is another way to look at the statement *the faster covers the longer distance*. It is about comparing the values of two sums. Consider this simple algebra:

$$\begin{array}{rcl} a & \leq & b \\ A & \leq & B \\ \hline a+A & \leq & b+B \end{array}$$

The rule applies even if we have more than just two terms:

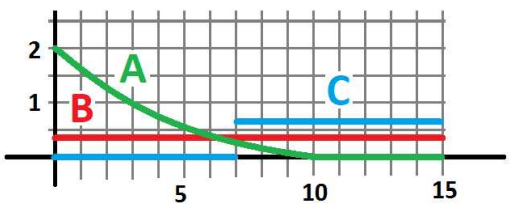
$$\begin{array}{rcl} a_m & \leq & b_m \\ a_{m+1} & \leq & b_{m+1} \\ \vdots & \vdots & \vdots \\ a_q & \leq & b_q \\ \hline a_m + \ldots + a_q & \leq & b_m + \ldots + b_q \end{array}$$

The summation is illustrated below:



Example 1.9.28: three runners, continued

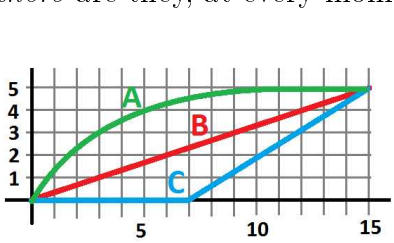
The graph shows the velocities of three runners in terms of time, n:

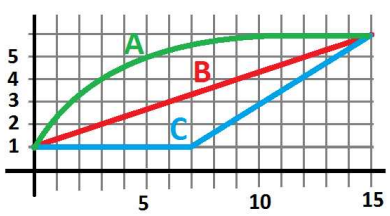


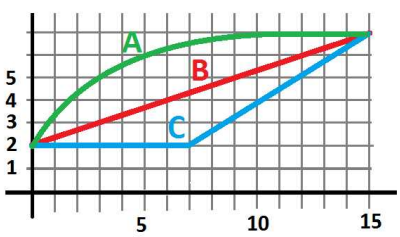
It's easy to describe *how* they are moving:

- A starts fast and the slows down.
- B maintains the same speed.
- C starts late and then runs fast.

But *where* are they, at every moment? There are several possible answers:

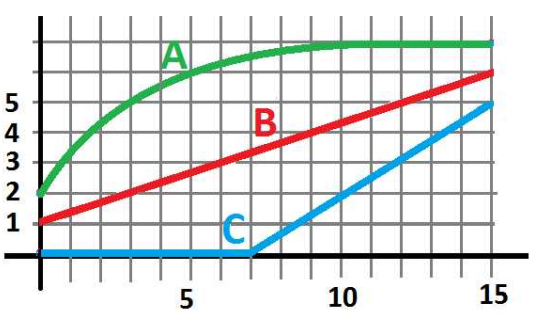






Which one is the right one depends on the starting point. Of course, a simple examination of the first graph doesn't prove that the three runners will arrive at the finish line at the same time.

Furthermore, if the requirement that they all start at the same location is lifted, the result will be different, for example:



Exercise 1.9.29

Suggest other graphs that match the description above.

Exercise 1.9.30

Plot the location and the velocity for the following trip: "I drove slowly, gradually speed up, stopped for a very short moment, and started but in the opposite direction, quickly accelerated, and from that point maintained the speed." Make up your own story and repeat the task.

Exercise 1.9.31

Draw a curve on a piece of paper, imagine that it represents your velocity, and then sketch what your locations would look like. Repeat.

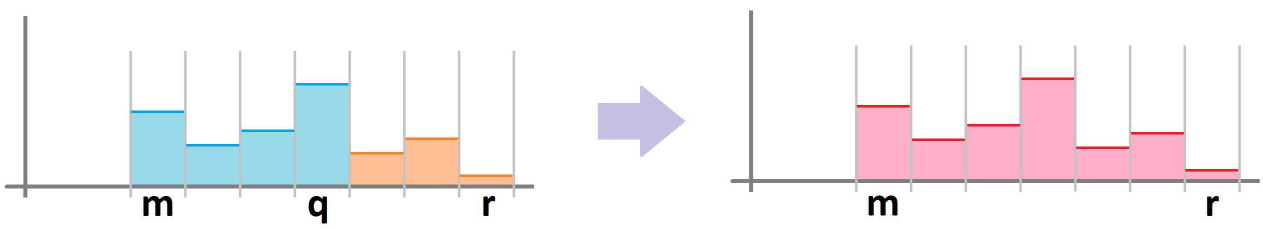
Here is another trivial statement about *motion*:

$$\begin{array}{ll} & \text{distance covered during the 1st hour} \\ + & \text{distance covered during the 2nd hour} \\ = & \text{distance during the two hours} \end{array}$$

The statement is about the fact that when adding, we can change the order of terms freely; this is called the *Associativity Property* of addition. At its simplest, it allows us to remove the parentheses:

$$\begin{array}{lllll} & (a_m + a_{m+1} + ... + a_{q-1} + a_q) & + & (a_{q+1} + a_{q+2} + ... + a_{r-1} + a_r) \\ = & a_m + a_{m+1} + ... + a_{q-1} + a_q & + & a_{q+1} + a_{q+2} + ... + a_{r-1} + a_r \\ = & a_m + a_{m+1} + & ... & +a_{r-1} + a_r \, . \end{array}$$

The three sums are shown below:



An abbreviated version of this identity is as follows.

Theorem 1.9.32: Additivity for Sums

The sum of the sums of two consecutive segments of a sequence is the sum of the combined segment.

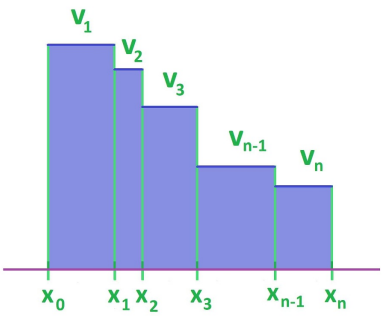
In other words, for any sequence a_n and for any m, q, r with $m \le q \le r$, we have:

$$\sum_{k=m}^{q} a_k + \sum_{k=q+1}^{r} a_k = \sum_{k=m}^{r} a_k$$

Example 1.9.33: Riemann sums of sequences

How do we deal with motion when the time moments aren't integers? What is the *displacement* then?

Suppose x_n is the sequence of locations and v_n the sequence of velocities:



Then their *Riemann sum* is defined to be the sequence of sums of the sequence of the product of v_n and the difference of x_n:

$$\sum_{k=1}^{n} v_n \, \Delta x_n \, .$$

We know from the last section that if y_n is the position, then the velocity is $v_n = \Delta y_n / \Delta x_n$. Therefore, the Riemann sum of the sequence of the velocity with respect to the sequence of time is the displacement. Of course, this is also the total area of the rectangles.

Example 1.9.34: sum of random sequence

We take a computer-generated random sequence (the values spread between -1 and 1):

n	a_n	sum
0		0.00
1	-0.76	-0.76
2	-0.42	-1.18
3	0.30	-0.87
4	-0.15	-1.02
5	-0.25	-1.27
6	-0.88	-2.15
7	0.41	-1.75
8	0.97	-0.78
9	-0.08	-0.85
10	-0.88	-1.74
11	-0.95	-2.69
12	0.49	-2.19
13	0.65	-1.54
14	0.65	-0.90
15	-0.13	-1.03
16	0.79	-0.24
17	-0.22	-0.46
18	0.20	-0.26
19	-0.09	-0.36
20	0.04	-0.32
21	0.17	-0.15
22	-0.78	-0.93



Then its sum is also random but might exhibit apparent periods of growth and decline ("streaks"). The numbers in the original sequence can represent the outcome of playing a hand of cards with the possibility of winning or losing an amount within \$1. Then the sequence of sums represents the amount the person has at every moment of time.

1.10. Sums of differences and differences of sums: motion

In this section, we start on the path of development of an idea that culminates with the cornerstone result of calculus, *the Fundamental Theorem of Calculus* (Chapter 3IC-1).

We know that *addition and subtraction undo each other*; it makes sense then that the operations of making the sequence of differences and making the sequence of sums will cancel each other too!

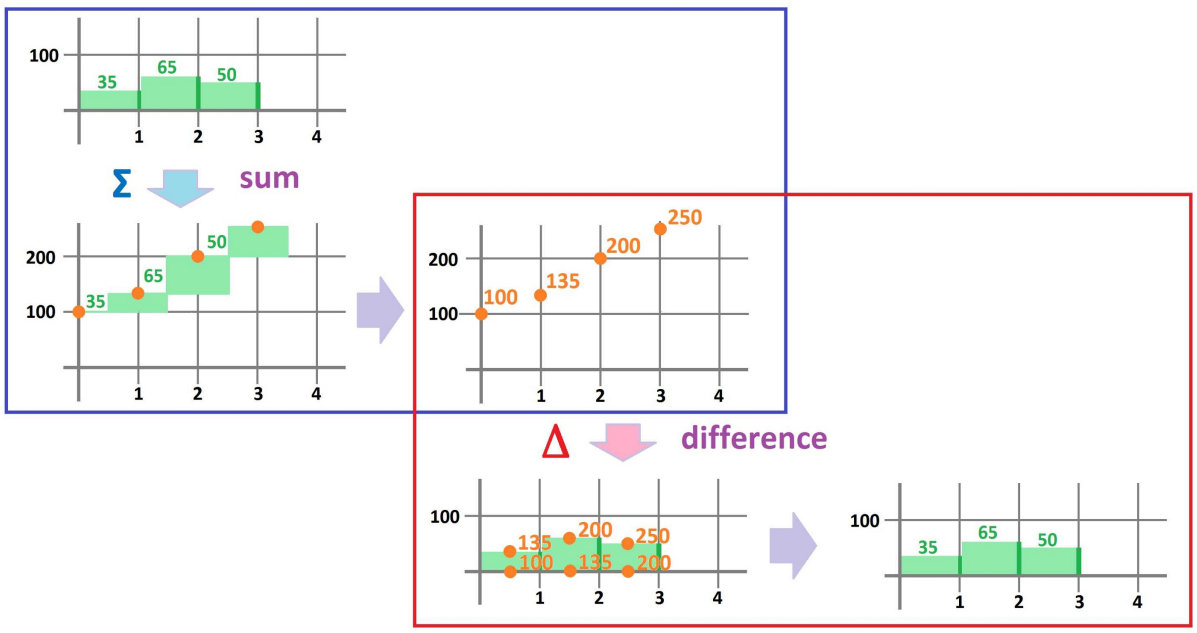
Example 1.10.1: broken odometer – broken speedometer

We know how to get the velocity from the location, and the location from the velocity. We expect that executing these two operations consecutively should bring us back where we started.

Let's take another look at the example of *two* computations about motion – a broken odometer and

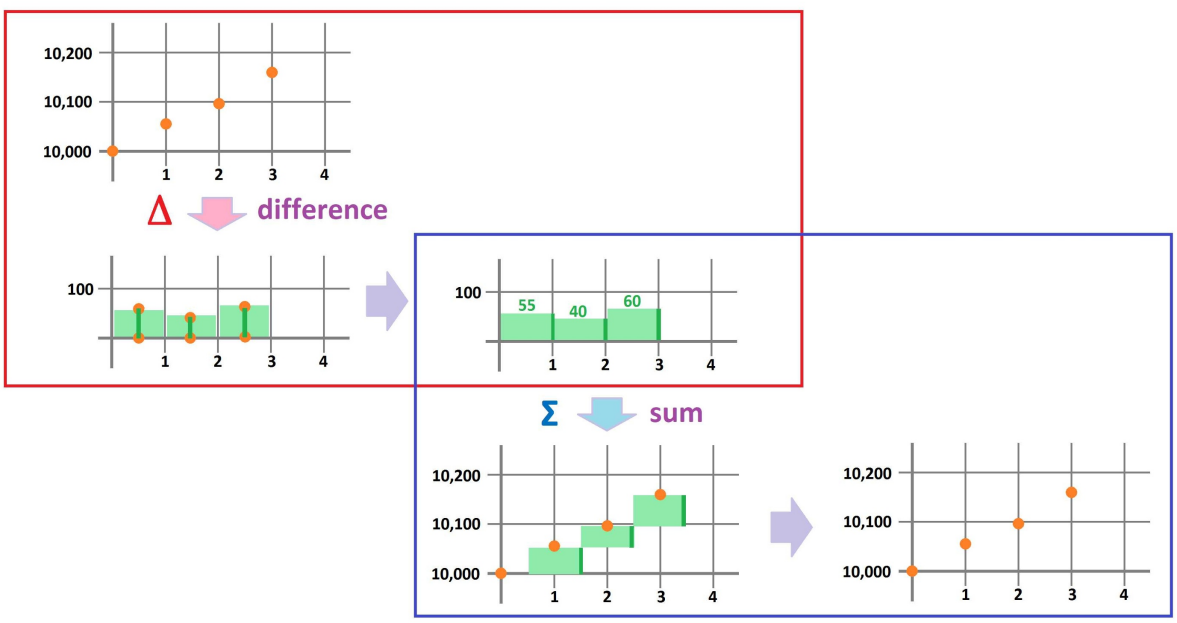
a broken speedometer – presented in the beginning of this chapter. The terminology has now been developed: Every time we speak of a sequence, we also speak of the sequence of its differences and the sequence of its sums.

In the first diagram, one first takes the velocity data and acquires the displacements via the sums, then someone else takes this displacement data and acquires the velocities by using the differences:



We are back to the original sequence.

In the second diagram, one first takes the location data and acquires the velocities via the differences, then someone else takes this velocity data and acquires the locations by using the sums:



We are back to the original sequence (provided we start at the same initial value).

Example 1.10.2: sequences given by lists

Below we have a sequence given by a list. We compute its sequence sums and then compute the

sequence of differences of the result:

a sequence:	3		1				0				-2				$\dots$
	$\downarrow$		$\downarrow$				$\downarrow$				$\downarrow$				$\dots$
its sums:	3														
	3	$+$	1	$=$	4										
					4	$+$	0	$=$	4	$\dots$					
									4	$+$	(-2)	$=$	2	$\dots$	
													2	$\dots$	
	$\downarrow$				$\downarrow$				$\downarrow$				$\downarrow$	$\dots$	
the sequence of sums:	3				4				4				2	$\dots$	
	$\searrow$		$\swarrow$		$\searrow$		$\swarrow$		$\searrow$		$\swarrow$			$\dots$	
the differences:		$4-3$				$4-4$				$2-4$				$\dots$	
		$\|\|$				$\|\|$				$\|\|$				$\dots$	
a new sequence:		1				0				-2				$\dots$	

We are back to the original sequence!

Exercise 1.10.3

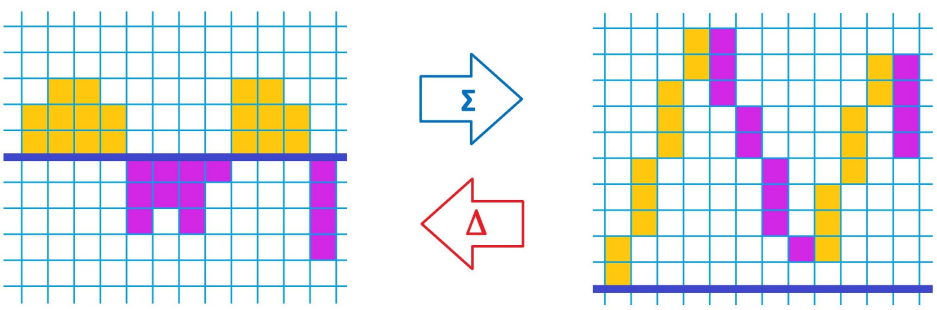
What happened to the very first term?

Exercise 1.10.4

Start with the sequence in the last example and use the diagrams to show that the sums of the differences give us the original sequence.

Example 1.10.5: sequences given by graphs

Just comparing the illustrations above demonstrates that the two operations – the difference and the sum – undo the effect of each other. The two operations are shown together below:



As you can see in the picture, the sum (left to right) stacks up the terms of the sequence on top of each other, while the difference (right to left) takes these apart.

Let's take care of the *algebra*.

These are the two facts we will be using:

1. Suppose we have a sequence, a_n. We compute its *difference*, a new sequence:

$$b_{n+1} = a_{n+1} - a_n \, .$$

2. Suppose we have a sequence, c_k. We compute its *sum*, a new sequence:

$$d_n = \sum_{k=1}^{n} c_k \, ,$$

or, recursively:

$$d_{n+1} = d_n + c_{n+1} \, .$$

We use this setup to answer the following two questions.

The first question we would like to answer is:

- ▶ *What is the difference of the sum?*

We start with c_n. Then, we have from (2) and (1), respectively:

$$d_{n+1} = d_n + c_{n+1} \quad \text{and} \quad b_{n+1} = d_{n+1} - d_n \,.$$

We substitute the first formula into the second (and then cancel):

$$b_{n+1} = d_{n+1} - d_n = (d_n + c_{n+1}) - d_n = c_{n+1} \,.$$

As we can see, the answer is:

- ▶ *The original sequence.*

The second question we would like to answer is:

- ▶ *What is the sum of the difference?*

We start with a_n. Then, we have from (1) and (2), respectively:

$$b_k = a_k - a_{k-1} \quad \text{and} \quad d_n = \sum_{k=1}^{n} b_k \,.$$

We substitute the first formula into the second (and then cancel):

$$d_n = \sum_{k=1}^{n} b_k = \sum_{k=1}^{n} (a_k - a_{k-1}) = (a_2 - a_1) + (a_3 - a_2) + (a_4 - a_3) + \ldots + (a_n - a_{n-1}) = -a_1 + a_n \,.$$

As we can see, the answer is:

- ▶ *The original sequence plus a number.*

We summarize these results in the form of the following two far-reaching theorems.

Theorem 1.10.6: Fundamental Theorem of Calculus of Sequences I

The difference of the sum of a sequence is that sequence; i.e., for all n, we have:

$$\Delta\left(\sum_{k=1}^{n} a_k\right) = a_n$$

The two operations *cancel* each other!

Theorem 1.10.7: Fundamental Theorem of Calculus of Sequences II

The sum of the difference of a sequence is that sequence plus a constant number; i.e., for all n, we have:

$$\sum_{k=1}^{n} \left(\Delta b_k\right) = b_n + C$$

The two operations – almost – cancel each other, again!

Example 1.10.8: fundamental theorems, computed

For larger sets of data, we use a spreadsheet. Recall the formulas:

- From a sequence to its sum:

```
=R[-1]C+RC[-1]
```

- From a sequence to its difference:

```
=RC[-1]-R[-1]C[-1]
```

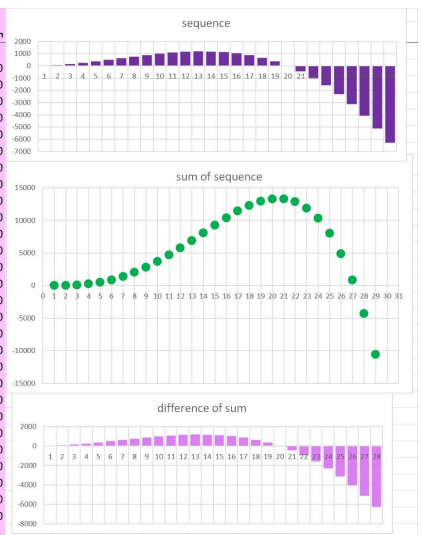
What if we combine the two consecutively? From a sequence to its difference to the sum of the latter:

n	a_n	$b_n=\Delta a_n$	Σb_n
0	0.0		0.0
1	49.0	49.0	49.0
2	192.0	143.0	192.0
3	423.0	231.0	423.0
4	736.0	313.0	736.0
5	1125.0	389.0	1125.0
6	1584.0	459.0	1584.0
7	2107.0	523.0	2107.0
8	2688.0	581.0	2688.0
9	3321.0	633.0	3321.0
10	4000.0	679.0	4000.0
11	4719.0	719.0	4719.0
12	5472.0	753.0	5472.0
13	6253.0	781.0	6253.0
14	7056.0	803.0	7056.0
15	7875.0	819.0	7875.0
16	8704.0	829.0	8704.0
17	9537.0	833.0	9537.0
18	10368.0	831.0	10368.0
19	11191.0	823.0	11191.0
20	12000.0	809.0	12000.0
21	12789.0	789.0	12789.0
22	13552.0	763.0	13552.0
23	14283.0	731.0	14283.0
24	14976.0	693.0	14976.0
25	15625.0	649.0	15625.0
26	16224.0	599.0	16224.0
27	16767.0	543.0	16767.0
28	17248.0	481.0	17248.0



It's the same curve! Now in the opposite order, from a sequence to its sum to the difference of the latter:

n	a_n	$b_n=\Sigma a_n$	Δb_n
0	0.0	0.0	
1	19.0	19.0	19.0
2	72.0	91.0	72.0
3	153.0	244.0	153.0
4	256.0	500.0	256.0
5	375.0	875.0	375.0
6	504.0	1379.0	504.0
7	637.0	2016.0	637.0
8	768.0	2784.0	768.0
9	891.0	3675.0	891.0
10	1000.0	4675.0	1000.0
11	1089.0	5764.0	1089.0
12	1152.0	6916.0	1152.0
13	1183.0	8099.0	1183.0
14	1176.0	9275.0	1176.0
15	1125.0	10400.0	1125.0
16	1024.0	11424.0	1024.0
17	867.0	12291.0	867.0
18	648.0	12939.0	648.0
19	361.0	13300.0	361.0
20	0.0	13300.0	0.0
21	-441.0	12859.0	-441.0
22	-968.0	11891.0	-968.0
23	-1587.0	10304.0	-1587.0
24	-2304.0	8000.0	-2304.0
25	-3125.0	4875.0	-3125.0
26	-4056.0	819.0	-4056.0
27	-5103.0	-4284.0	-5103.0
28	-6272.0	-10556.0	-6272.0



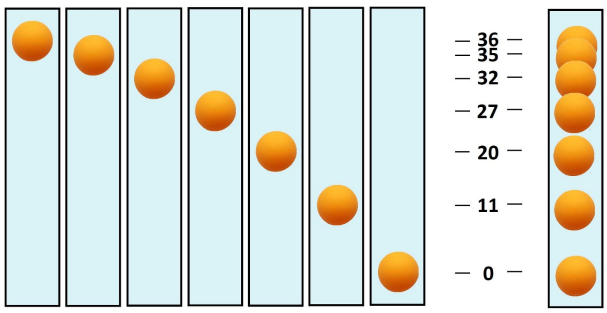
It's the same curve!

Exercise 1.10.9

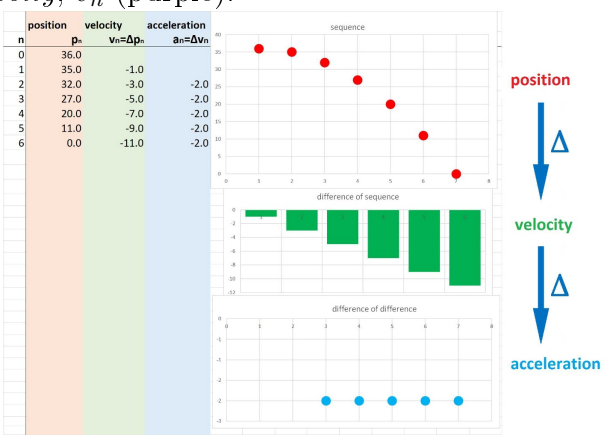
What would the resulting curve look like if we started at another point?

Example 1.10.10: falling ball, acceleration

In an example from earlier in this chapter, we viewed the experimental data of the heights of a ping-pong ball falling down:



Just as before, we use a spreadsheet to plot the *location* sequence, p_n (green). We then compute the difference of p_n, i.e., the *velocity*, v_n (purple):



n	position p_n	velocity $v_n=\Delta p_n$	acceleration $a_n=\Delta v_n$
0	36.0		
1	35.0	-1.0	
2	32.0	-3.0	-2.0
3	27.0	-5.0	-2.0
4	20.0	-7.0	-2.0
5	11.0	-9.0	-2.0
6	0.0	-11.0	-2.0

It looks like a straight line. But this time, we take one more step: We compute the difference of the velocity sequence. It is the *acceleration*, a_n (blue). It appears constant! There might be a law of nature here.

Example 1.10.11: shooting a cannon

Let's accept the premise put forward in the last example, that *the acceleration of free fall is constant*. Then we can try to predict the behavior of an object shot in the air – from any initial height and with any initial velocity.

The direction of our computation is opposite to that of the last example: We assume that we know the acceleration, then derive the velocity, and then derive the location (altitude) of the object in time. While we used *differences* in the last example, we use *sums* now:



n	acceleration a_n	velocity $v_n=\Sigma a_n$	position $p_n=\Sigma v_n$
0	-9.8	200.0	100.0
1	-9.8	190.2	290.2
2	-9.8	180.4	470.6
3	-9.8	170.6	641.2
4	-9.8	160.8	802.0
5	-9.8	151.0	953.0
6	-9.8	141.2	1094.2
7	-9.8	131.4	1225.6
8	-9.8	121.6	1347.2
9	-9.8	111.8	1459.0
10	-9.8	102.0	1561.0
11	-9.8	92.2	1653.2
12	-9.8	82.4	1735.6
13	-9.8	72.6	1808.2
14	-9.8	62.8	1871.0
15	-9.8	53.0	1924.0
16	-9.8	43.2	1967.2
17	-9.8	33.4	2000.6
18	-9.8	23.6	2024.2
19	-9.8	13.8	2038.0
20	-9.8	4.0	2042.0
21	-9.8	-5.8	2036.2
22	-9.8	-15.6	2020.6
23	-9.8	-25.4	1995.2
24	-9.8	-35.2	1960.0
25	-9.8	-45.0	1915.0
26	-9.8	-54.8	1860.2
27	-9.8	-64.6	1795.6
28	-9.8	-74.4	1721.2
29	-9.8	-84.2	1637.0
30	-9.8	-94.0	1543.0
31	-9.8	-103.8	1439.2

Above we show a projectile launched from a 100-meter tall building vertically up in the air with a speed of 100 meters per second (the gravity causes acceleration of -9.8 meters per second squared).

We can see that it will reach its highest point in about 20 seconds and will hit the ground in about 40 seconds.

Exercise 1.10.12

How high does the projectile go in the above example?

Exercise 1.10.13

Using the above example, how long will it take for the projectile to reach the ground if fired *down*?

Exercise 1.10.14

Use the above model to determine how long it will take for an object to reach the ground if it is dropped. Make up your own questions about the situation and answer them. Repeat.

Exercise 1.10.15

Suppose the time moments are given by another sequence (an arithmetic progression). Compute the velocity and the acceleration from the table below:

	time	height
n	t_n	a_n
1	.00	36
2	.05	35
3	.10	32
4	.15	25
5	.20	20
6	.25	11
7	.30	0

This study of motion continues throughout the book.

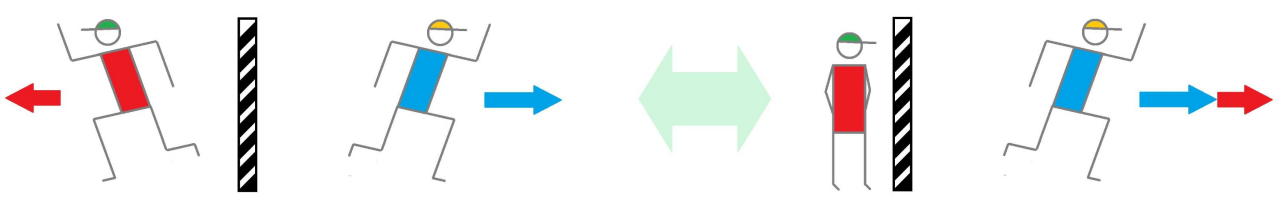
1.11. The algebra of sums and differences

What happens to their *differences* and their *sums* after an algebraic operation is carried out with a pair of sequences? There are a few shortcut properties.

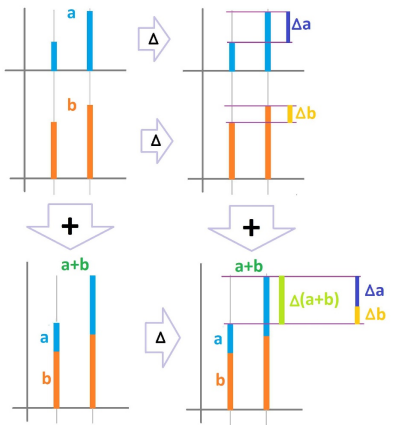
Here is an elementary statement about *motion*:

> ▶ `IF` two runners are running away from a post, `THEN` their relative velocity is the sum of their respective velocities.

It's as if the one runner is standing still while the other is running with the combined speed:



The idea why we *add* their differences when we add sequences is illustrated below:



Here, the bars that represent the change of the values of the sequence are stacked on top of each other. The heights are then added to each other, and so are the height differences. The algebra behind this geometry is very simple:

$$(A+B)-(a+b)=(A-a)+(B-b)\,.$$

It's the *Associative Rule* of addition.

The idea above is equally applicable to runners who change how fast they run; we speak of sequences:

Theorem 1.11.1: Sum Rule for Differences

The difference of the sum of two sequences is the sum of their differences.

In other words, for any two sequences a_n, b_n, their sequences of differences satisfy:

$$\Delta(a_n+b_n)=\Delta a_n+\Delta b_n$$

Proof.

$$\begin{aligned}\Delta(a_n+b_n) &= (a_{n+1}+b_{n+1})-(a_n+b_n)\\ &= (a_{n+1}-a_n)+(b_{n+1}-b_n)\\ &= \Delta a_n+\Delta b_n\,.\end{aligned}$$

Example 1.11.2: difference of sum

Consider the sum of an arithmetic progression and a geometric progression:

$$\Delta(a+mn+ar^n)=\Delta(a+mn)+\Delta(ar^n)=m+ar^n(r-1)\,.$$

Now, differences and sums are matched up according to the *Fundamental Theorems of Calculus of Sequences* presented in the last section! It should be no surprise then that there is a matching theorem about sums.

When two sequences are added to each other, what happens to their sums? This simple algebra, the *Associative Property* combined with the *Commutative Property*, tells the whole story:

re-arrange the sum of four numbers:

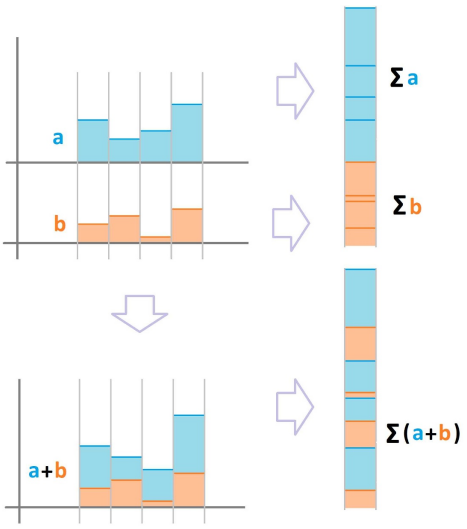
	a	$+$	b	$=$	$(a+b)$,
	$+$		$+$		$+$
	A	$+$	B	$=$	$(A+B)$
$=$	$(a+A)$	$+$	$(b+B)$	$=$	$a+A+b+B$

The rule applies even if we have more than just two terms; it's about re-arranging terms of sequences:

re-arrange the sum of two sequences:

a_p	$+$	b_p	$=$	$(a_p+b_p)+$	
a_{p+1}	$+$	b_{p+1}	$=$	$(a_{p+1}+b_{p+1})+$	
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
a_q	$+$	b_q	$=$	(a_q+b_q)	
$=(a_p+\ldots+a_q)$	$+$	$(b_p+\ldots+b_q)$	$=$	$(a_p+b_p)+$	$\ldots$ $+(a_q+b_q)$

The summation is illustrated below:



An abbreviated version of this formula is as follows.

Theorem 1.11.3: Sum Rule for Sums

The sum of the sums of two sequences is the sum of the sequence of the sums.

In other words, if a_n and b_n are sequences, then, for any p, q with $p \leq q$, their sequences of sums satisfy:

$$\sum_{n=p}^{q} a_n + \sum_{n=p}^{q} b_n = \sum_{n=p}^{q} (a_n + b_n)$$

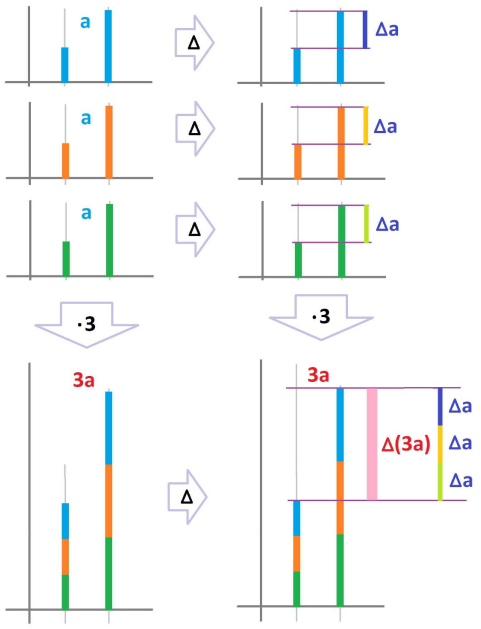
Exercise 1.11.4

Derive this theorem from the last one, reverse.

Here is another simple statement about *motion*:

> ▶ `IF` the distance is re-scaled, such as from miles to kilometers, `THEN` so is the velocity – at the same proportion.

The idea why a *proportional* change causes the same proportional change in the differences is illustrated below (tripling):



Here, if the heights triple, then so do the height differences. The algebra behind this geometry is very simple:

$$kA - ka = k(A - a)\,.$$

It's the *Distributive Rule*. This is how it applies to sequences.

Theorem 1.11.5: Constant Multiple Rule for Differences

The difference of a multiple of a sequence is the multiple of the sequence's difference.

In other words, for any sequence a_n, the sequence of differences satisfies:

$$\Delta(ka_n) = k\Delta a_n$$

Proof.

$$\begin{aligned}\Delta(ka_n) &= ka_{n+1}k - ka_n \\ &= ka_{n+1}k - ka_n \\ &= k\Delta a_n\,.\end{aligned}$$

The theorem can also be interpreted as follows: If the distances are proportionally increased, then so are the velocities needed to cover them, in the same period of time.

Is there a matching statement about sums? Yes, but let's first look at motion again: If your velocity is tripled, then so is the distance you have covered.

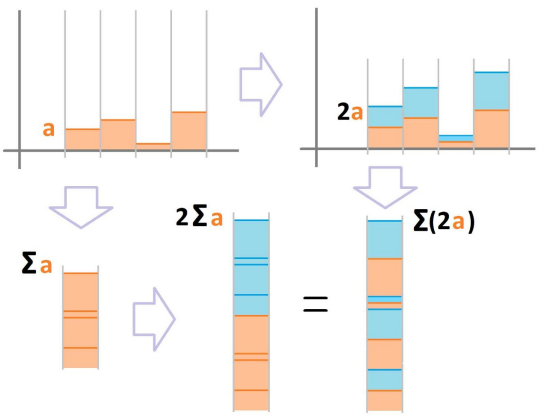
When a sequence is multiplied by a constant, what happens to its sums? This simple algebra, the *Distributive Property*, tells the whole story:

$$\text{take out a common factor:} \quad \begin{array}{rcccc} k & \cdot & (& a & + & b) \\ & = & & ka & + & kb \end{array}$$

The rule applies even if we have more than just two terms; it's about factoring:

$$\text{take out a common factor:} \quad \begin{array}{rcl|c|l} k & \cdot & a_p & = & k \cdot a_p + \\ k & \cdot & a_{p+1} & = & k \cdot a_{p+1} + \\ \vdots & \vdots & \vdots & & \vdots \\ k & \cdot & a_q & = & k \cdot a_q \\ \hline = k \cdot a_p + & \ldots & + k \cdot a_q & = & k \cdot (a_p + \ldots + a_q) \end{array}$$

This summation is illustrated below:



An abbreviated version of this formula is as follows.

Theorem 1.11.6: Constant Multiple Rule for Sums

The sum of a multiple of a sequence is the multiple of its sum.

In other words, if a_n is a sequence, then for any p, q with $p \leq q$ and any real k, its sequence of sums satisfies:

$$\sum_{n=p}^{q}(ka_n) = k\sum_{n=p}^{q} a_n$$

Exercise 1.11.7

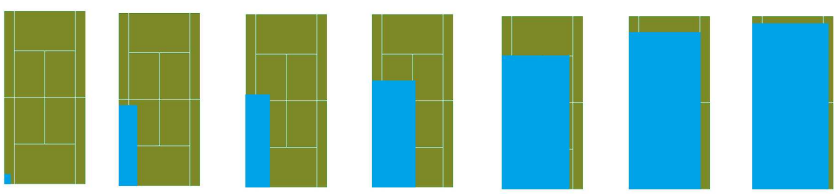
Derive this theorem from the last one, reverse.

Now we go beyond addition and multiplication by a constant.

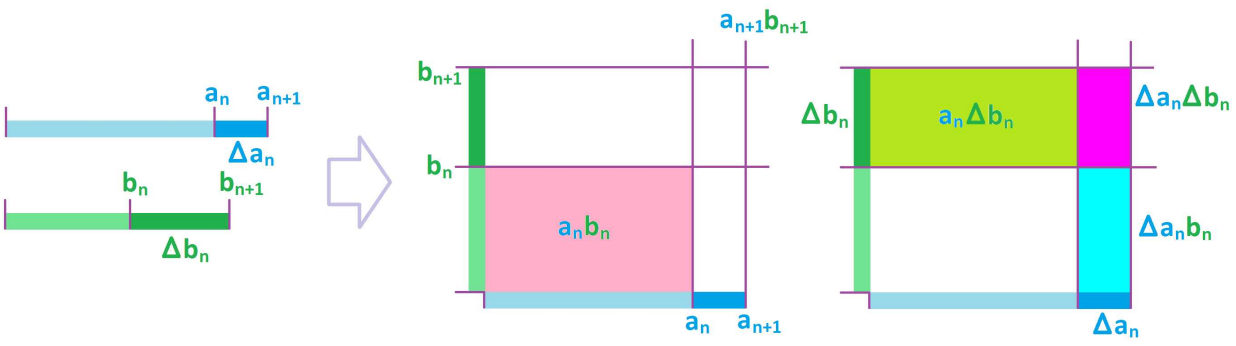
Let's imagine this:

> ▶ If two groups of runners are unfolding a tarp (or unfurling a flag) while running east and north, respectively, what is happening to the area of this rectangle?

They may be running at different speeds:



Then, the product of two sequences is interpreted as the *areas* of the rectangles formed by the sequences, illustrated below:



As the width and the depth are increasing, so is the area of the rectangle. We can see that the increase of the area cannot be expressed entirely in terms of the increases of the width and depth! This increase is split into two parts corresponding to the two terms in the right-hand side of the formula below.

Theorem 1.11.8: Product Rule for Differences

The difference of the product of two sequences is found as a combination of these sequences and either of the two differences.

Specifically, for any two sequences a_n, b_n, the sequence of differences of the product satisfies:

$$\Delta(a_n \cdot b_n) = a_{n+1} \cdot \Delta b_n + \Delta a_n \cdot b_n$$

Proof.

$$\begin{aligned}
\Delta(a_n \cdot b_n) &= a_{n+1} \cdot b_{n+1} - a_n \cdot b_n && \text{Insert terms.}\\
&= a_{n+1} \cdot b_{n+1} - a_{n+1} \cdot b_n + a_{n+1} \cdot b_n - a_n \cdot b_n && \text{Factor.}\\
&= a_{n+1} \cdot (b_{n+1}) - b_n) + (a_{n+1} - a_n) \cdot b_n\\
&= a_{n+1} \cdot \Delta b + \Delta a_n \cdot b_n \,.
\end{aligned}$$

Example 1.11.9: difference of product

Consider the product of an arithmetic progression and a geometric progressions:

$$\Delta(mn \cdot ar^n) = \Delta(mn) \cdot ar^n + mn\Delta(ar^n) = mar^n + mnar^n(r-1)\,.$$

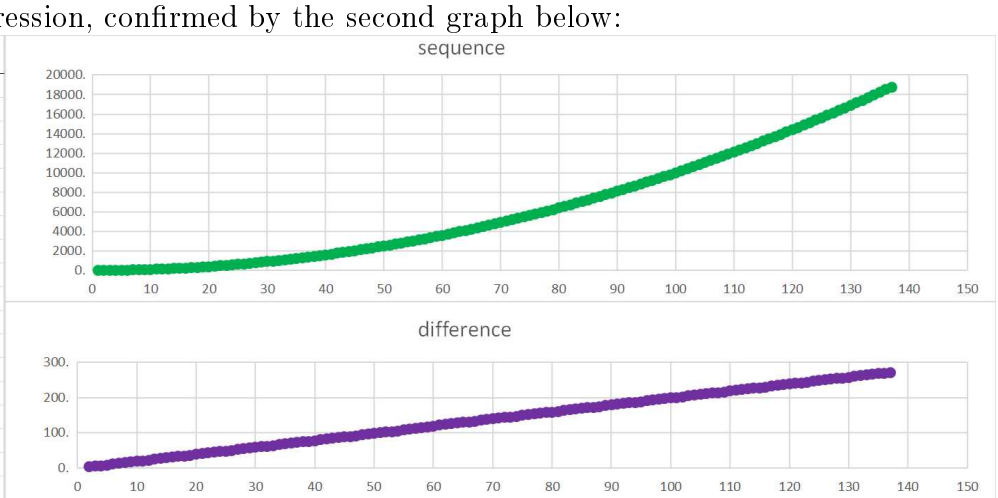
Example 1.11.10: difference of square

The difference of the square sequence $a_n = n^2$ is computed with the *Product Rule*:

$$\Delta(n^2) = \Delta(n \cdot n) = (n+1) \cdot \Delta(n) + \Delta(n) \cdot (n) = (n+1) + (n) = 2n + 1\,.$$

It's an arithmetic progression, confirmed by the second graph below:

n	$a_n=n^2$	Δa_n
1	1.	
2	4.	3.
3	9.	5.
4	16.	7.
5	25.	9.
6	36.	11.
7	49.	13.
8	64.	15.
9	81.	17.
10	100.	19.
11	121.	21.
12	144.	23.
13	169.	25.
14	196.	27.
15	225.	29.
16	256.	31.
17	289.	33.
18	324.	35.



Exercise 1.11.11

Find a formula for the difference of the power sequence, $a_n = n^m$, using the Binomial Theorem.

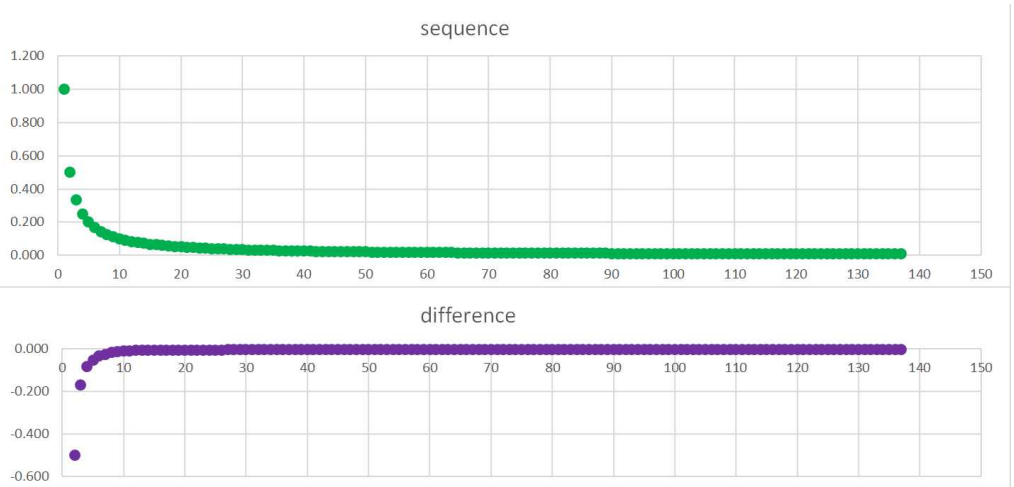
Example 1.11.12: difference of sequence of reciprocals

Let's find the formula for the difference of the reciprocals:

$$\Delta\left(\frac{1}{n}\right) = \frac{1}{n+1} - \frac{1}{n} = \frac{n-(n+1)}{n(n+1)} = -\frac{1}{n(n+1)}\,.$$

The sequence decreases but slower and slower:

n	$a_n=1/n$	Δa_n
1	1.000	
2	0.500	-0.500
3	0.333	-0.167
4	0.250	-0.083
5	0.200	-0.050
6	0.167	-0.033
7	0.143	-0.024
8	0.125	-0.018
9	0.111	-0.014
10	0.100	-0.011
11	0.091	-0.009
12	0.083	-0.008
13	0.077	-0.006
14	0.071	-0.005
15	0.067	-0.005
16	0.063	-0.004
17	0.059	-0.004
18	0.056	-0.003



Theorem 1.11.13: Quotient Rule for Differences

The difference of the quotient of two sequences is found as a combination of these sequences and either of the two differences.

Specifically, for any two sequences a_n, b_n, the sequence of differences of the quotient satisfies:

$$\Delta\left(\frac{a_n}{b_n}\right) = \frac{a_{n+1}\cdot\Delta b_n + \Delta a_n \cdot b_n}{b_n b_{n+1}}$$

Exercise 1.11.14

Prove the theorem.

There are no matching statements for *sums* as simple as these two.

Here precalculus as a *preview of calculus* ends, and precalculus as a *prerequisite for calculus* starts.

Chapter 2: Sets and functions

Contents

2.1 Sets and relations 98
2.2 Functions 106
2.3 Sequences are numerical functions 112
2.4 How numerical functions emerge: optimization 117
2.5 Set building 124
2.6 The xy-plane: where graphs live... 132
2.7 Linear relations 138
2.8 Relations vs. functions 142
2.9 A function as a black box 148
2.10 Give the function a domain... 158
2.11 The graph of a function 165
2.12 Linear functions 171
2.13 Algebra creates functions 179
2.14 The image: the range of values of a function 192

2.1. Sets and relations

In mathematics, we refer to any loose collection of objects or entities – of any nature – as a *set*.

For example, is this a circle of marbles that we see in a bag? No, the marbles it is made of aren't connected to each other or to any location. One shake and the circle is gone:

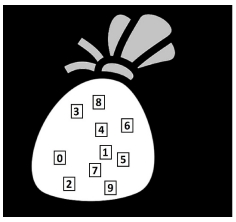


Example 2.1.1: sets as lists

Sets given explicitly – as lists – are the simplest ones:

- A roster of students: Adams, Adkins, Arrows, ...
- A list of numbers: 1, 2, 3, 4, ...
- A list of planets: Mercury, Venus, Earth, Mars, ...

The order at which they appear on the list is not a part of the information we care about when we speak of sets. Here is a bag of numbers:



Example 2.1.2: "sets"

The idea of set contrasts with such expressions as "a set of silverware" when the word "set" suggests a certain structure: specific types of knives and forks with a specific place in the box. It is the same set, mathematically, whether the items are arranged in a box or piled up on the counter. A set of encyclopedia consists of books that can be arranged alphabetically or chronologically or randomly.

Warning!

Even though we try to provide a precise definition of every new concept, the idea of set is so general that we will have to rely on examples.

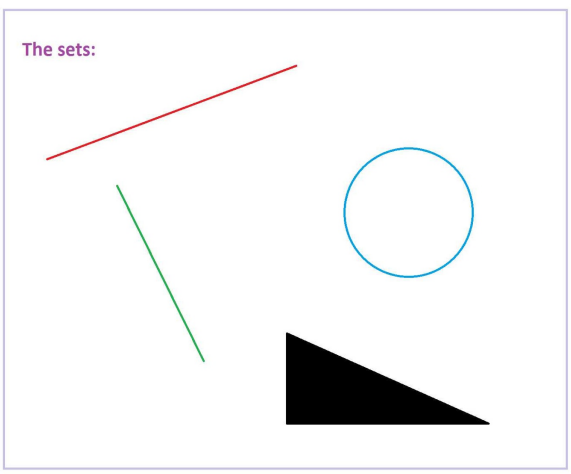
What creates a set is our knowledge or ability to determine whether an object *belongs or does not belong* to it. A list is one such possibility. Another is a condition to be verified.

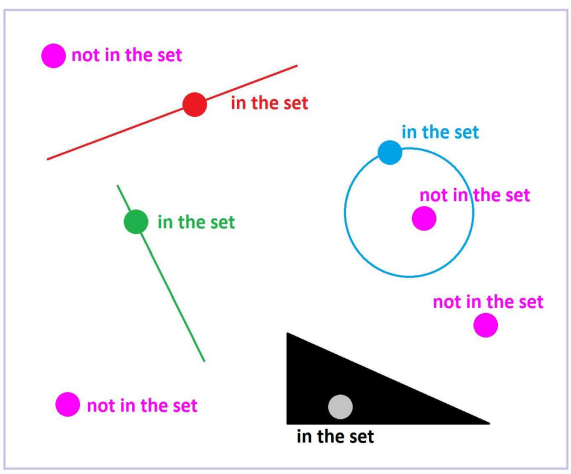
Example 2.1.3: sets via conditions

A roster of a class produces a set of the students in this class. It's a list! On the other hand, the *female* students in the class also form a set even if there is no such list; we can just go down the roster and determine if a student belongs to this new set. Similarly, the students with an A on the last test also – implicitly – form a set.

Example 2.1.4: sets in math

A lot of sets examined early in this book will be sets of *numbers*. For example, take the set of *even numbers*; then we know that 2 belongs to it but 3 does not. We simply check the condition: Is the number divisible by 2? Another example from familiar parts of mathematics is sets of *points* on the plane: straight lines, triangles, circles and other curves, etc.:





We can always tell whether a point belongs to the set!

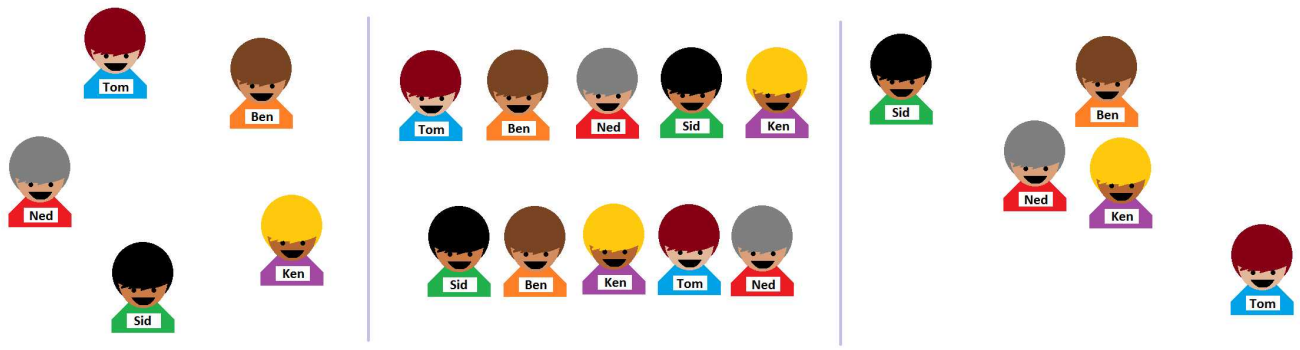
Example 2.1.5: non-sets

If the condition is vague, we don't have a set: "interesting novels", "bad paintings", etc. When the condition is nonsensical, we don't have a set either: "fast trees", "blue numbers", etc.

Exercise 2.1.6

Give your own examples of (a) sets as lists, (b) sets defined via conditions, and (c) non-sets.

In the rest of this chapter we will use the following example. These *five boys* form a set:



On the one hand, they are individuals and can always be told from each other. On the other hand, they are unrelated to each other: We can list them in any order, we can arrange them in a circle, a square, or at random; we can change the distances between them, and so on. It's the same set! The members of a set are called its *elements*.

Our set is nothing but a *list*:

- Tom
- Ken
- Sid
- Ned
- Ben

Or: "Tom, Ken, Sid, Ned, Ben", in any order.

Warning!

As there is no order, the *elements* of a set aren't to be confused with the *terms* of a sequence as the latter are ordered.

There is a specific mathematical notation for finite sets; we put the list in *braces*:

List notation for sets

$$\{A, B, C, D\}$$

It reads "the set with elements A, B, C, D".

All of these are equally valid representations of our set:

$$\begin{aligned} & \{ \text{ Tom , Ken , Sid , Ned , Ben } \} \\ = \ & \{ \text{ Ned , Ken , Tom , Ben , Sid } \} \\ = \ & \{ \text{ Ben , Ken , Sid , Tom , Ned } \} \\ = \ & \ldots \end{aligned}$$

Exercise 2.1.7

How many such representations are there? Hint: In how many ways can you permute these five elements?

Just as the boys have names, the set also needs one. We can call this set "Team", or "Boys", etc. To keep things compact, let's give it a short name, say X:

$$X = \{ \text{Tom} , \text{Ken} , \text{Sid} , \text{Ned} , \text{Ben} \}.$$

We say then that Tom (Ken, etc.) is an element of set X, as well as:

- Tom *belongs* to X, or
- X *contains* Tom.

Just as we want to be clear when two numbers are equal, we want the same clarity for sets. The following will be assumed to be known.

Definition 2.1.8: equal sets

Two sets X and Y are said to be *equal* to each other if:
- every element of X is also an element of Y, and conversely,
- every element of Y is also an element of X.

Repetitions aren't allowed! Or, at least, they are to be eliminated:

$$\{ \text{Tom} , \text{Ken} , \text{Sid} , \text{Ned} , \text{Ben} , \text{Ben} \} \xrightarrow{\text{remove repetitions!}} \{ \text{Tom} , \text{Ken} , \text{Sid} , \text{Ned} , \text{Ben} \}$$

It's the same set!

We can form other sets from the same elements. We can combine those five elements into any set with any number of elements as long as there is no repetition; for example, we can create these new sets:

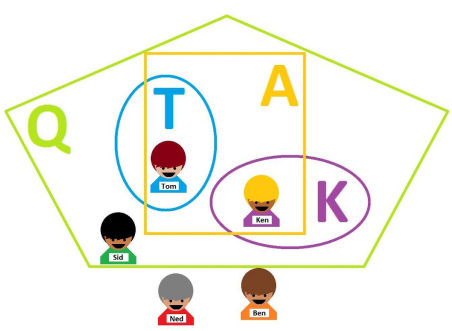
$$T = \{ \text{Tom} \}, \quad K = \{ \text{Ken} \}, \quad S = \{ \text{Sid} \}, \quad N = \{ \text{Ned} \}, \quad \ldots$$
$$A = \{ \text{Tom} , \text{Ken} \}, \quad B = \{ \text{Sid} , \text{Ned} \}, \quad \ldots$$
$$Q = \{ \text{Tom} , \text{Ken} , \text{Sid} \}, \quad \ldots$$

The following will be routinely used.

Definition 2.1.9: subset

A set A is called a *subset* of a set X if every element of A is also an element of X.

This is how we mark subsets when the set is shown:



Exercise 2.1.10

How many subsets of 3 elements does the set have? Hint: In how many ways can you choose three elements out of five?

We will use the following notation to convey that idea:

Subset

$$A \subset X$$

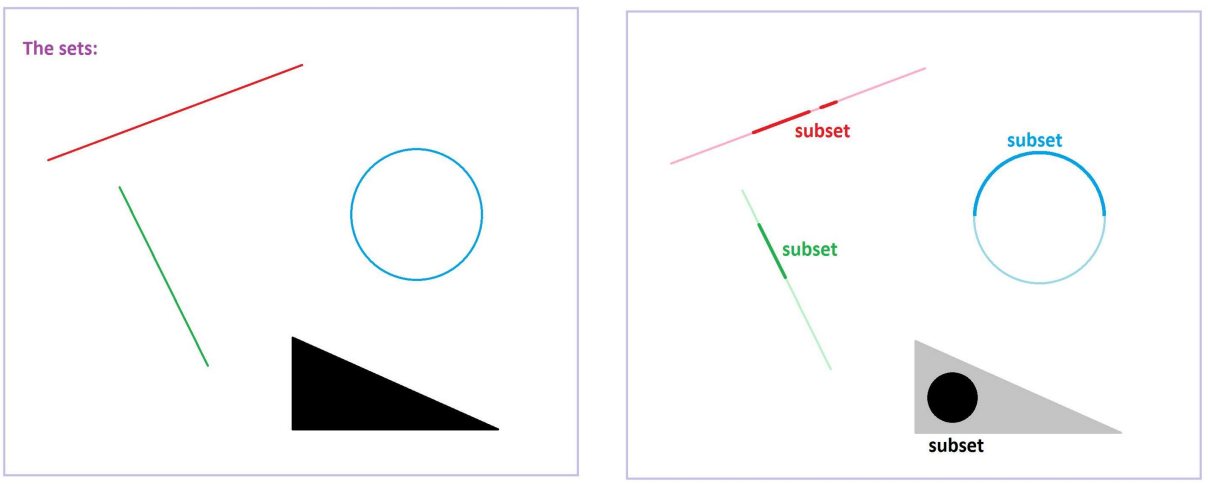
The notation resembles the one for numbers: $1 < 2$, $3 < 5$, etc. Indeed, a subset is, in a sense, "smaller" than the set that contains it.

Warning!

A subset doesn't *have to* be literally smaller because a set is a subset of itself. Furthermore, an infinite set might have a subset just as infinite...

Example 2.1.11: plane shapes

We see subsets of geometric figures in the plane:

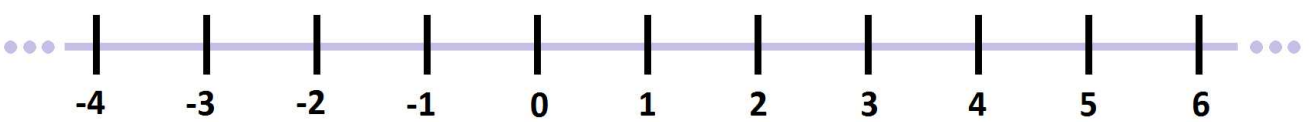


In Chapter 1, we started our study of numbers with the following two sets. We chose to speak of locations spaced over an infinite straight line associated with the *integers*, denoted by

$$\mathbf{Z} = \{..., -3, -2, -1, 0, 1, 2, 3, ...\},$$

and we spoke of time moments spaced over the *natural numbers*:

$$\mathbf{N} = \{0, 1, 2, 3, ...\}.$$



In the meantime, the set of real numbers is denoted by $\mathbf{R}$. It is visualized as the x-axis:



We can express the relation between these sets using the new notation:

$$\mathbf{N} \subset \mathbf{Z} \subset \mathbf{R}.$$

Exercise 2.1.12

Is a sequence a set?

To continue with our example, suppose there is *another*, unrelated, set, say Y, the set of these four balls:

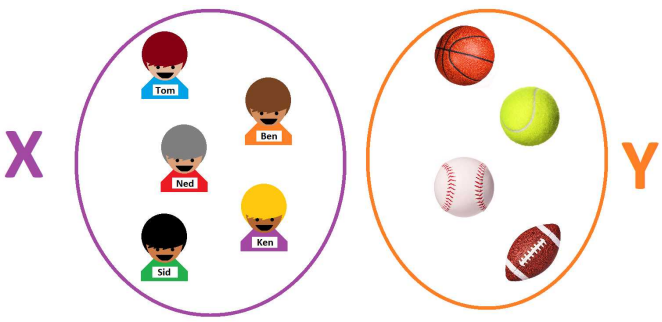


Just as X, set Y has no structure. Just as X, it's just a list:

$$\begin{aligned} Y &= \{ \text{ basketball , tennis , baseball , football } \} \\ &= \{ \text{ football , baseball , tennis , basketball } \} \\ &= \ldots \end{aligned}$$

We can remove balls from the set, creating subsets of Y.

Now, let's put the two sets, X and Y, next to each other and ask ourselves: Are these two sets related to each other somehow?



Yes, boys like to play sports! Let's make this idea specific. Each boy may be *interested in a particular sport* or he may not. For example, suppose this is what we know:

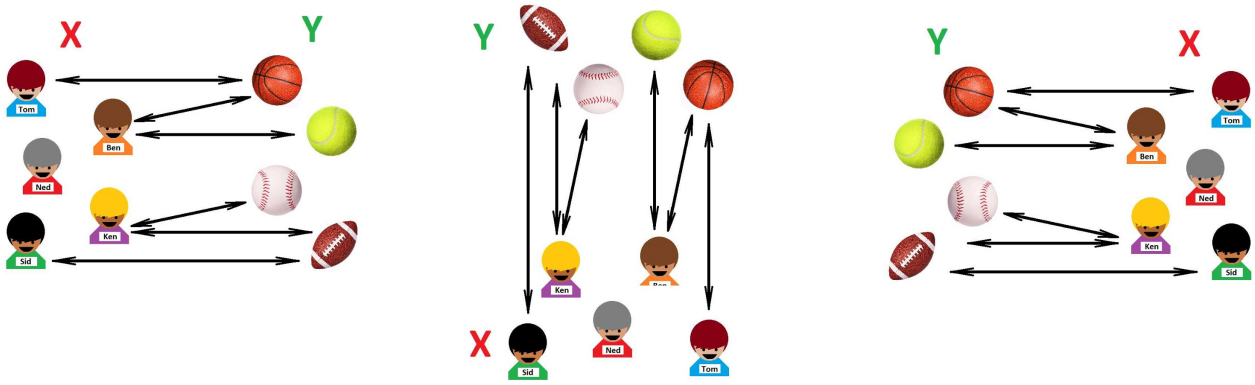
- Tom likes basketball.
- Ben likes basketball and tennis.
- Ken likes baseball and football.
- Ben likes football.

And that's all each likes.

As a result, we have the following:

▶ An element of set X is *related* to an element of set Y.

In order to visualize these relations, let's connect each boy with the corresponding ball by a line segment with arrows at the ends, while the two sets may be placed arbitrarily against each other:



This visualization helps us discover that Ned doesn't like sports at all. As you can see, this is a two-sided correspondence: Neither of the two elements at the ends of the line comes first or second. The same applies to the sets: Neither of the two sets comes first or second. In fact, we can derive these new facts about the preferences either from the original list or from the image on the right:

- Basketball is liked by Tom and Ben.
- Tennis is liked by Ben.
- Baseball is liked by Ken.
- Football is liked by Ken and Sid.

We have, therefore, a *list of pairs*:

- Tom & basketball
- Ben & basketball
- Ben & tennis
- Ken & baseball
- Ken & football
- Ben & football

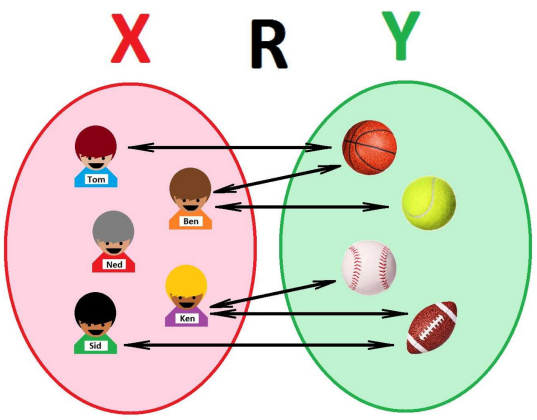
The following concept will be commonly used throughout.

Definition 2.1.13: relation between sets

Any set of pairs (x, y), with x taken from a set X and y from a set Y, is called a *relation* between sets X and Y.

In other words, this is a set of arrows.

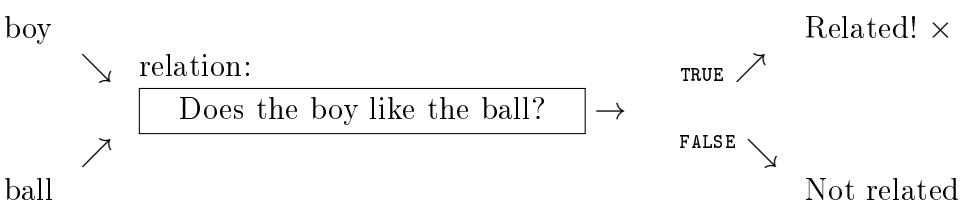
There may be many different relations between a pair sets; let's call this one R:



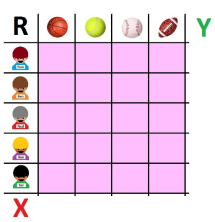
Warning!

We don't require every element to have a corresponding element in the other set.

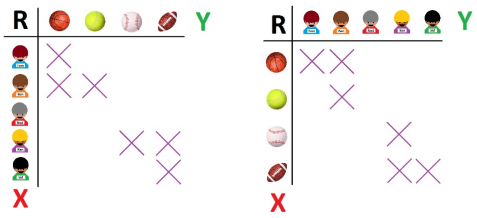
We can also represent the relation by the following diagram:



When the sets are lists, relations are *tables*. Let's make a table for R! We put the boys in the leftmost column and the balls in the top row. There are 20 cells:



If the boy likes the sport, we put a mark in the boy's row and the ball's column (left):



Or, we can put the boys in the first row and the balls in the first column (right)! In other words, we flip the table about its *diagonal*. These are two visualizations of the same relation.

This is what such a visualization looks like when we use a *spreadsheet* instead:

X\Y	basketball	tennis	baseball	football
Tom	x			
Ben	x	x		
Ned				
Ken			x	x
Sid				x

Exercise 2.1.14

Based on the relation R presented above, create a new relation called, say, S, that relates the boys and the sports they *don't* like. Give an arrow representations and a table representations of S.

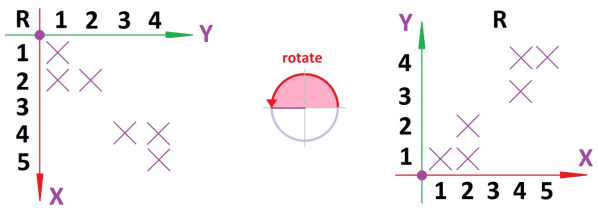
Exercise 2.1.15

Are there any relations on the subsets of the two sets?

Any collection (a set) of marks in such a table creates a relation, and conversely, a relation is nothing but a collection of marks in this table.

Throughout the early part of this book, we will concentrate on sets that consist of *numbers*. Even though the set of numbers does have a structure (Chapter 1), the ideas presented above still apply.

To illustrate these ideas, how about we simply *rename* the boys as numbers, $1-5$? And we rename the balls as numbers too, $1-4$. Then the former belong to the set of real numbers, while the latter belong to another copy of this set. We then draw these number lines along the sides of the table (seen on the left):



These axes are also labeled to avoid confusion between the two very different sets. Furthermore, on the right, the table is rotated (90 degrees counterclockwise). This table is then called the *graph* of the relation. The placements of the two sets within the tables can still be interchanged.

Exercise 2.1.16

When the rows and the columns are interchanged, is there anything that is preserved?

Exercise 2.1.17

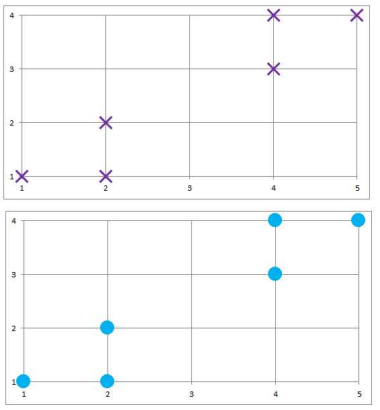
Finish the sentence: "This renaming of the boys is a ___."

Suppose we have the elements of the sets renamed as numbers (left), then we capture the relation as a list of pairs of elements of X and Y (middle), and finally, the graph of the relation can be plotted automatically

by the spreadsheets:

X\Y	basketball	tennis	baseball	football
Tom	x			
Ben	x	x		
Ned				
Ken			x	x
Sid				x

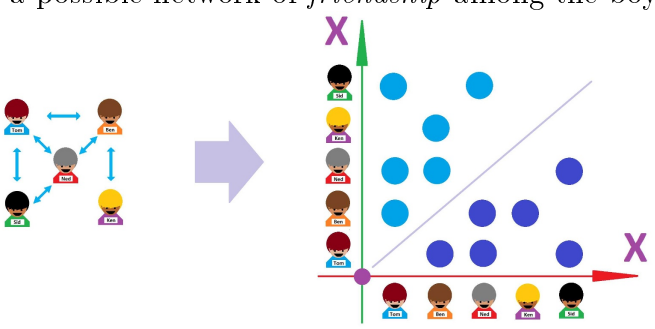
x	y
1	1
2	1
2	2
4	3
4	4
5	4



It is called a "scatter chart".

Example 2.1.18: networks as relations

The plot below represents a possible network of *friendship* among the boys:



We can still represent this as a relation; we just choose the sets X and Y to be the same. The nature of this kind of relation isn't just two-sided; it's "symmetric": If Tom is a friend of Ben, then Ben is also a friend of Tom, and vice versa. That is why each arrow in the diagram on the left is represented by *two* marks in the graph of the relation on the right.

Exercise 2.1.19

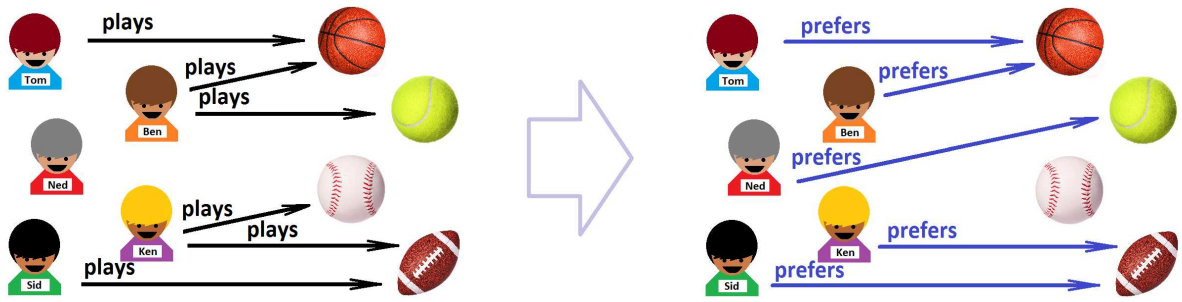
If the five boys decided to have a ping-pong tournament, what relation does it create on X?

2.2. Functions

Let's go back to our running example and change the question from:

- "What sports has the boy played today?" to:
- "Which sport does the boy *prefer* to play?"

The idea is that everyone, even Ned, has a preference and exactly one. This is the transition:

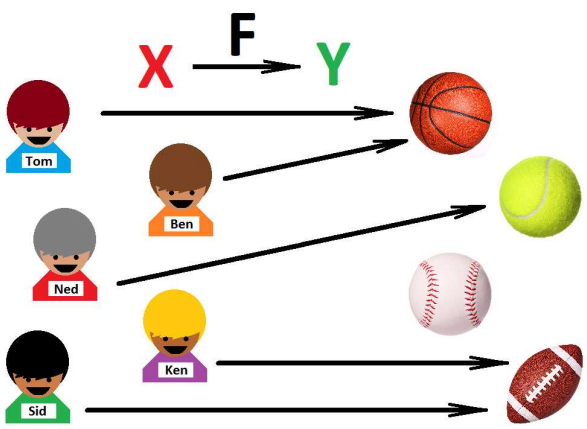


To make sense of this new setup, we had to erase one of the two arrows that start at Ben and one of the two arrows that start at Ken and we had to add an arrow for Ned.

In a relation, the two sets involved play equal roles. Instead, we now take the point of view of the boys. We will explore a new relation:

- Tom prefers basketball.
- Ben prefers basketball.
- Ned prefers tennis.
- Ken prefers football.
- Sid prefers football.

We move from our two-ended arrows (or line segments) to regular arrows:



We maintain the rule: there is exactly one y for each x. As a result, the equality between the two sets is gone: X comes first, Y second.

This is a special kind of relation called a *function*; let's call this one F. What makes it special is that there is exactly one ball for each boy. Below is the common notation:

Function from set to set

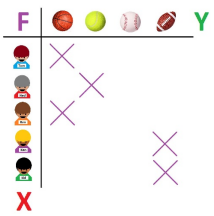
$$F : X \to Y$$

or

$$X \xrightarrow{F} Y$$

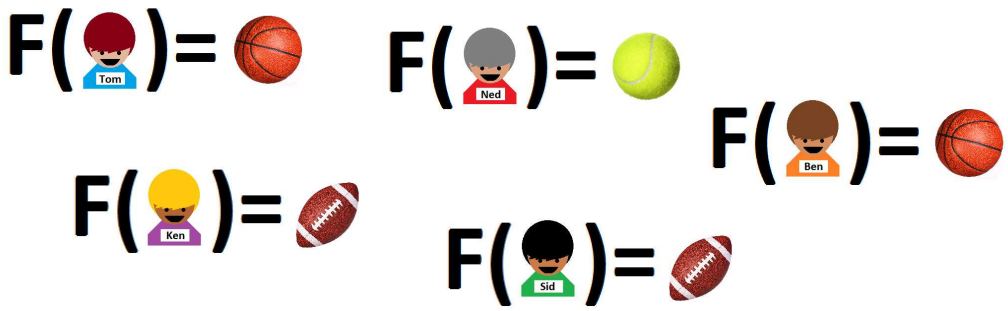
It reads "function F from X to Y".

Each element of X has only one arrow originating from it. Then, the table of this kind of relation must have exactly one mark in each row:



Our function F is a *procedure* that answers the question: *Which ball does this boy prefer to play with?* In fact, it answers *all* these questions! Conversely, a function is nothing but these answers. Each arrow clearly identifies the *input* – an element of X – of this procedure by its beginning and the *output* – an element of Y – by its end.

Each arrow in the diagram of F corresponds to a row of the table (and vice versa). The information contained in each is more commonly written in an *algebraic* manner, as follows:



The function F is then a question-answering machine: If you input the name of the boy, it will produce the name of the ball he prefers as the output.

This is the notation for the output of a function F when the input is x:

Input and output of function

$$F(x) = y$$

or

$$F : x \mapsto y$$

It reads: "F of x is y".

In other words, we have

$$F(\text{ input }) = \text{ output}$$

and

$$F : \text{ input } \mapsto \text{ output.}$$

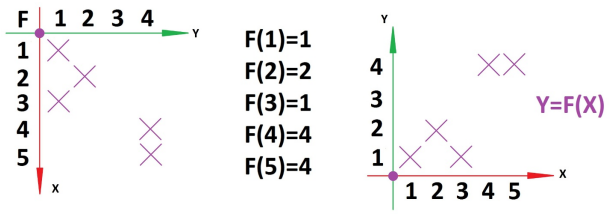
Just like any relation, a function can be represented in full by providing a list of pairs, x and y. This time, it's the *list of all inputs and their outputs*:

$$\begin{aligned} F(\text{ Tom }) &= \text{basketball} \\ F(\text{ Ned }) &= \text{tennis} \\ F(\text{ Ben }) &= \text{basketball} \\ F(\text{ Ken }) &= \text{football} \\ F(\text{ Sid }) &= \text{football} \end{aligned}$$

This notation will be, by far, the most common way of representing functions.

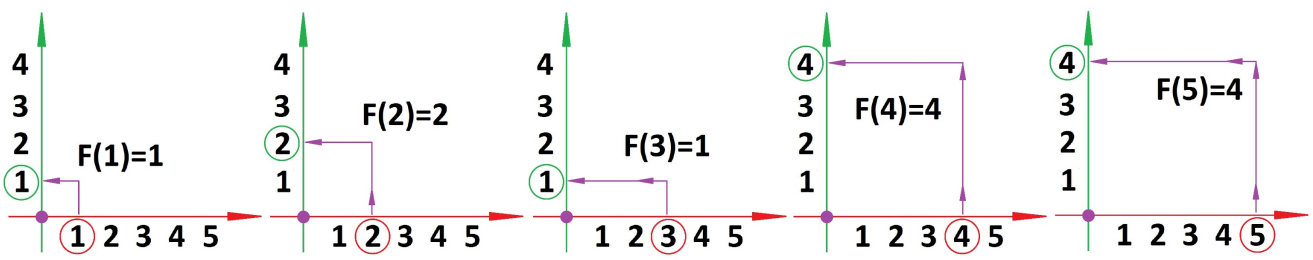
Throughout the early part of this book, we will concentrate on functions the inputs and the outputs of which are *numbers*.

To illustrate this idea, let's again *rename* the boys as numbers, $1-5$, and rename the balls as numbers, $1-4$. The table of our relation takes this form (seen on the left):



What makes the table of a *function* special is that it must have exactly one mark in each *column*. The values of F have also been rewritten (center). We then rotate the table counterclockwise (right) because it is traditional to have the inputs along a horizontal line (left to right) and the outputs along a vertical line (bottom to top).

The latter table is called the *graph* of the function. The arrows are still there:



Exercise 2.2.1

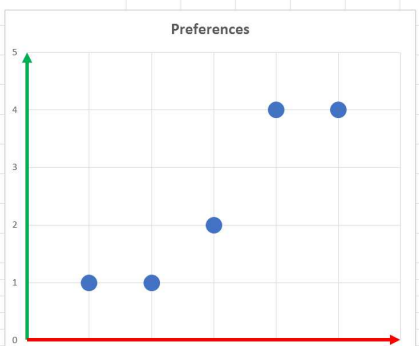
Finish the sentence: "This renaming of the boys (and the balls) is also a ___."

We can put the data, as before, in a *spreadsheet* and then plot it automatically:

X\Y	basketball	tennis	baseball	football
Tom	x			
Ben	x			
Ned		x		
Ken				x
Sid				x

X\Y	1	2	3	4
1	x			
2	x			
3		x		
4				x
5				x

X	Y
1	1
2	1
3	2
4	4
5	4



There is only one cross in every row!

Example 2.2.2: relations and function in spreadsheets

Here is an example of how common spreadsheets are discovered to contain relations and functions. Below, we have a list of faculty members in the first column and a list of faculty committees in the first row. A cross mark indicates on which committee this faculty member sits, while "C" stands for "chair":

	Committees:				
Faculty:	Promotion	Research	Hiring	Teaching	Grants
Ayanna Mcmanis	x		C	x	
Angila Mcgillivray		x		x	
Letitia Hollmann		x		x	C
Nancee Loftin					
Vannesa Labonte					
Caterina Mcclaren					
Leone Marmon	C	x			x
Marguerita Lucien				x	
Alyce Leininger				x	
Youlanda Suter				x	
Jeniffer Carey		x			
Noel Nokes		C	x		
Lory Ledger				x	
Whitney Lanigan					
Loretta Vicente	x			C	
Estela Lowery	x				
Lauri Ormsby	x				x
Zella Treiber					x
Bruno Reichman				x	
Delma Saeger		x	x		

The table gives a relation between these sets: $X = \{$ faculty $\}$ and $Y = \{$ committees $\}$. However, this is not a function. But there is a function $F : Y \to X$ indicating the chair of the committee.

Exercise 2.2.3

Think of other functions present in the spreadsheet.

Exercise 2.2.4

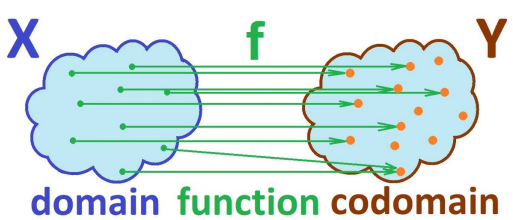
Suggest functions in the situation when an employer maintains a list of employees, with each person identified as a member of one of the projects.

Exercise 2.2.5

What functions do you see below?

Name	Nationality	Position	Goals
Arthur Albiston	Scotland	FB	7
Anderson	Brazil	MF	9
John Aston Jr.	England	FW	27
John Aston Sr.	England	FB	30
Gary Bailey	England	GK	0
Tommy Bamford	Wales	FW	57
Frank Barrett	Scotland	GK	0
Frank Barson	England	HB	4
Fabien Barthez	France	GK	0
Bobby Beale	England	GK	0
David Beckham	England	MF	85
Alex Bell	Scotland	HB	10
Ray Bennion	Wales	HB	3
Dimitar Berbatov	Bulgaria	FW	56
Henning Berg	Norway	DF	3
Johnny Berry	England	FW	45
George Best	Northern Ireland	FW	179

A common way to visualize the concept of set – especially when the sets cannot be represented by mere lists – is to draw a shapeless blob in order to suggest the absence of any internal structure or relation between the elements:



A common way to visualize the idea of a function between such sets is to draw a few arrows.

The new concept is central to our study.

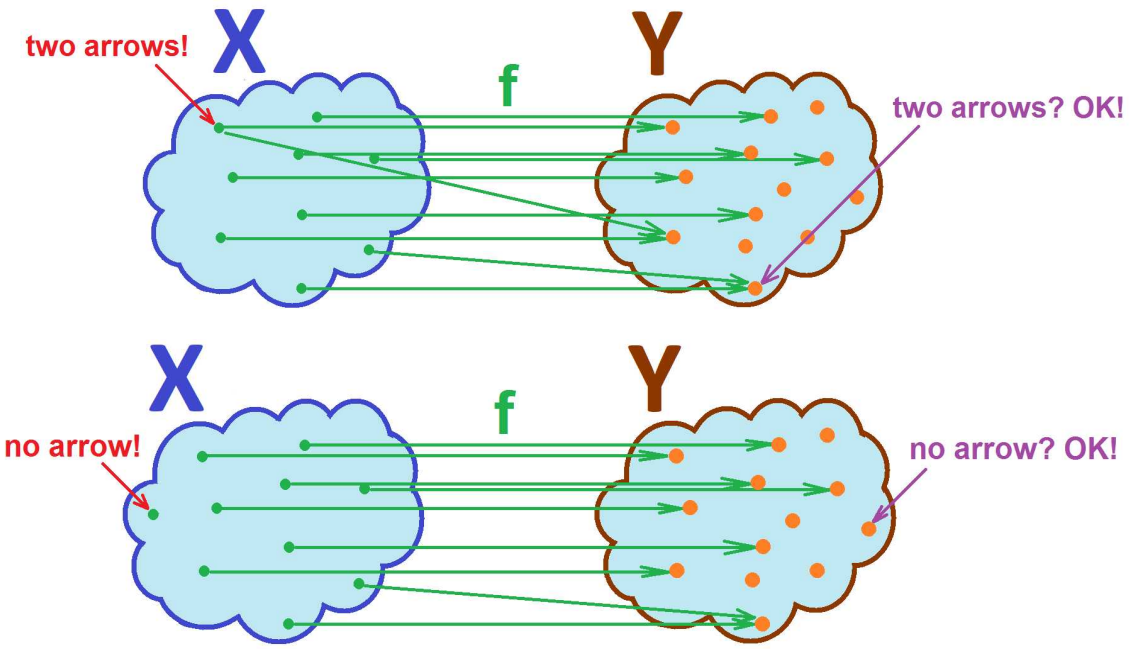
Definition 2.2.6: function

A *function* is a rule or procedure f that assigns to any element x in a set X, called the *input set* or the *domain* of f, exactly one element y, which is then denoted by

$$y = f(x)\,,$$

in another set Y. The latter set is called the *output set* or the *codomain* of f. The inputs are collectively called the *independent variable*; the outputs are collectively called the *dependent variable*. We also say that *the value of x under f* is y.

This definition fails for a relation that has too few or too many arrows for a given x. Below, we illustrate how the requirement may be violated, in the domain (left):



These are *not* functions. Meanwhile, we also see what shouldn't be regarded as violations, in the codomain (right).

Below is the summary of the main construction in this section.

Theorem 2.2.7: When Relation Is Function

Suppose X and Y are sets and R is a relation between X and Y. Then:

- *Relation R represents some function F from X to Y, $F : X \to Y$, if and only if for each x in X there is exactly one y in Y such that x and y are related by R.*
- *Relation R represents some function G from Y to X, $G : Y \to X$, if and only if for each y in Y there is exactly one x in X such that x and y are related by R.*

Exercise 2.2.8

Split either part into a statement and its converse.

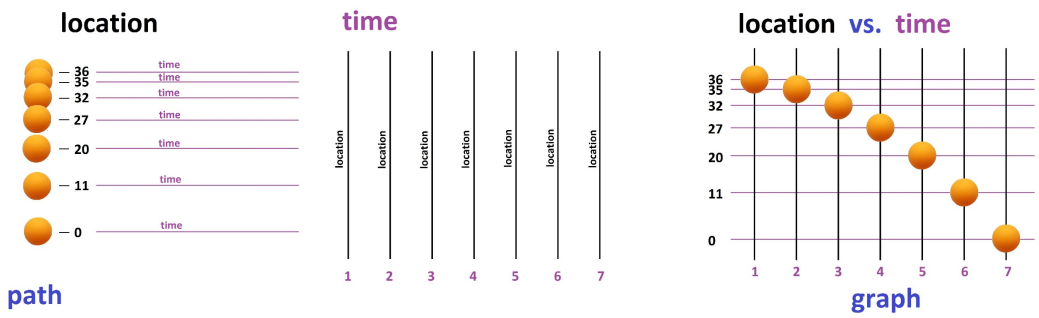
When our sets are sets of numbers, the relations are often given by *formulas*. In that case, the above issue is resolved with algebra.

Exercise 2.2.9

What function can you think of from the set X of the boys to the sets of: letters, numbers, colors, geographic locations? Think of others.

2.3. Sequences are numerical functions

In Chapter 1, we visualized a sequence of positions of a falling ball by "separating space and time". We gave the former a real number line and the latter a line of integers:



But the latter is also a subset of the real numbers:

$$\{1, 2, 3, 4, 5, 6, 7\} \subset \mathbf{R}\,.$$

We have, therefore, a function. In fact, the sequence was represented as a list of pairs of inputs and outputs, just as any function:

time	location
n	a_n
1	36
2	35
3	32
4	27
5	20
6	11
7	0
X	Y

So, sequences are simply functions with integer inputs and real number outputs.

Definition 2.3.1: sequence as function

A *sequence* is a function from a subset of the integers, $A \subset \mathbf{Z}$, to the real numbers:

$$a : A \to \mathbf{R}\,.$$

The following will be our main interest.

Definition 2.3.2: numerical function

A *numerical function* is a function from a subset of the real numbers, $X \subset \mathbf{R}$, to the real numbers:

$$f : X \to \mathbf{R}\,.$$

Since every subset of the integers is also a subset of the reals,

$$A \subset \mathbf{Z} \subset \mathbf{R},$$

every sequence is also a numerical function. However, not every numerical function is a sequence! Simply put,

$$a_{1/2} \text{ doesn't make sense.}$$

Let's compare:

- A sequence: The input variable is n, an integer; the output variable is $y = a_n$, a real number.
- A numerical function: The input variable is x, a real number; the output variable is $y = f(x)$, another real number.

We compare the notation too, side by side:

Function vs. sequence

name of the function

$f(x)$ vs. a_n

name of the input variable

value of the input variable

$f(3) = 5$ vs. $a_3 = 5$

value of the output variable

But is this transition – to more a complex structure – worth it?

If X is the set of time moments and Y is the set of locations on the road, we can see a way to study *motion*! Indeed, a function $F : X \to Y$ answers the question:

▶ At this moment of time, x, what is the location, y, where we are?

This is a *function* because we can't be at two locations at the same time!

Example 2.3.3: functions arise from motion

Here is a very simple example: Suppose we move to the next milestone every minute for 2 minutes, starting at the 0 location on the real line. To make this more precise, we may ask:

▶ At time x, which milestone $y = F(x)$ did we see last?

Then the *list* of values of F appears:

time, X	locations, Y
first moment	first milestone
second moment	second milestone
third moment	third milestone

With a sequence, we would choose the *domain* to be

$$X = \{0,\ 1,\ 2\}.$$

and the *codomain* to be

$$Y = \mathbf{R}.$$

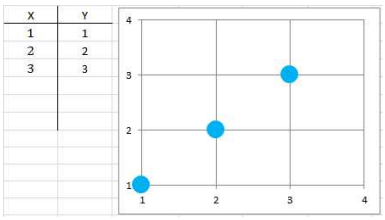
Then our list and the table are as follows:

time, X	locations, Y
1	1
2	2
3	3

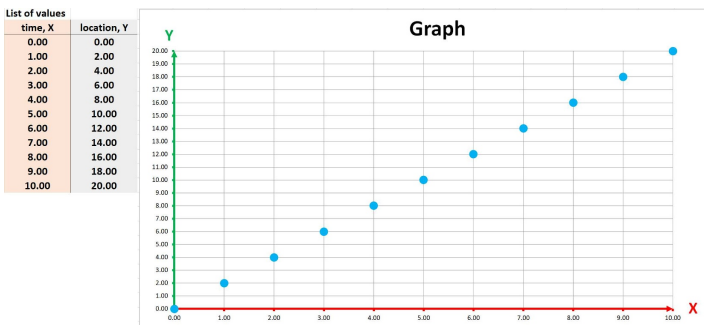
and

time location	1	2	3
1	×		
2		×	
3			×

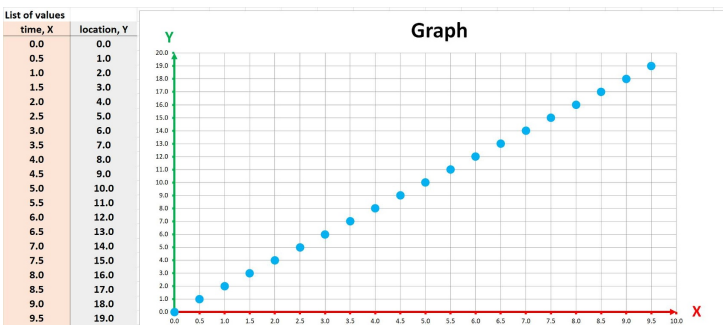
Finally, this is the *graph* of F:



Driving at a constant speed, we progress 2 miles every minute. We produce more data with a spreadsheet:



The graph might give an impression that we skip milestones! Then, in order to capture our motion more thoroughly, we simply start asking the same question: *At time x, which milestone $y = F(x)$ did we see last?* but every 30 seconds. We introduce half-minute marks to our set of inputs:



Is this still a sequence? Yes, if n measures half-minutes. With a function, we simply keep the set of outputs $Y = \mathbf{R}$ and change the set of inputs X from (every mile):

$$X = \{0,\ 1,\ 2,\ 3,\ 4,\ 5,\ ...,\ 9\}$$

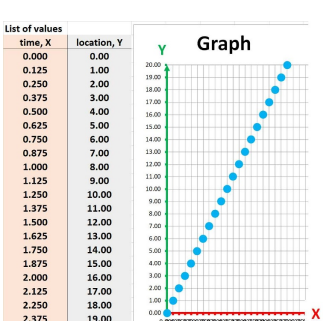
to (every 1/2-mile):

$$X = \{0,\ .5,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ ...,\ 8.5,\ 9,\ 9.5,\ 10\}\,.$$

The problem is solved – until we choose to drive even faster. Driving 4 miles per minute will require the outputs to be (every 1/4-mile):

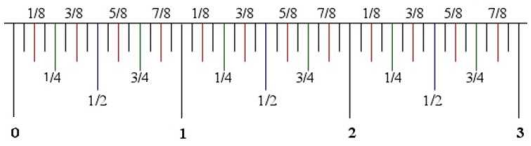
$$X = \{0, .25,\ .5,\ .75,\ 1,\ 1.25,\ 1.5,\ ...\,, 9.75,\ 10\}\,.$$

And so on:



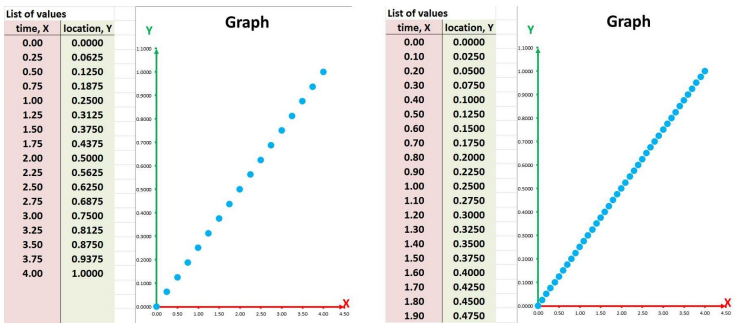
Accommodating finer and finer representations of space or time will require us to continue to divide the real

line. It starts to look like a *ruler*:



We commonly think of motion as a continuous progress through physical space. This requires us to think of both space and time as infinitely divisible.

But how do we visualize such functions? We still represent them as sequences of pairs of numbers – and then plot their graphs – but with a clear understanding that some of the inputs are missing.



We insert more inputs as necessary. When there are enough of them, they start to form a curve!

Exercise 2.3.4

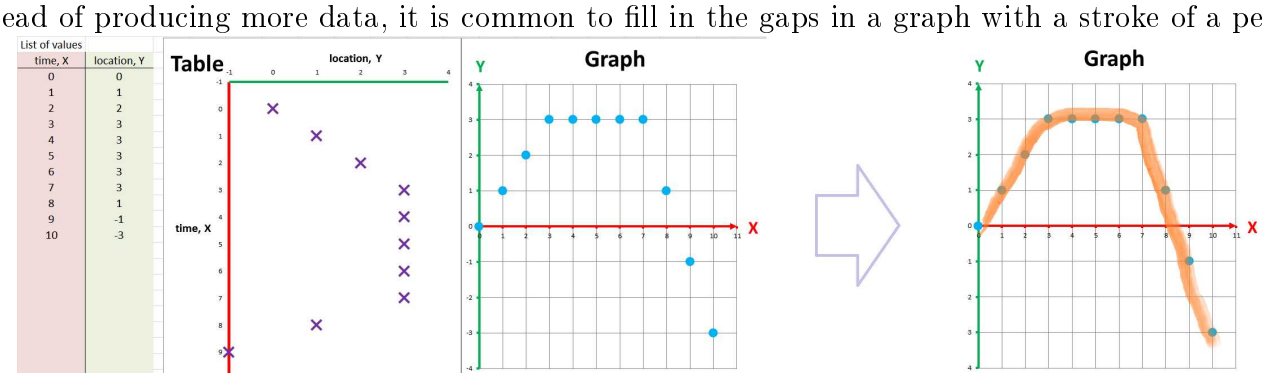
A car starts moving west from town A at a constant speed of 40 miles an hour. Town B is located 50 miles east of A. Represent the distance from town B to the car as a function of time.

Exercise 2.3.5

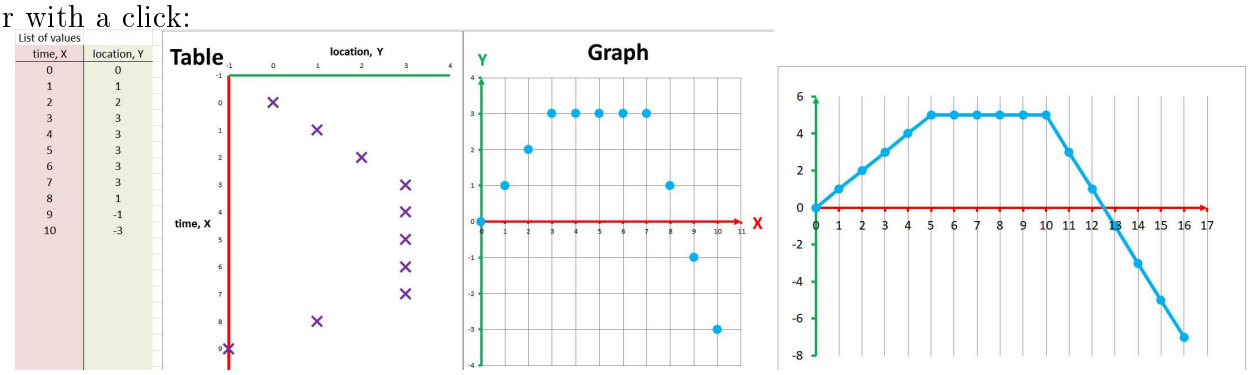
A car starts moving west from town A at a constant speed of 40 miles an hour. At the same time, another car starts moving west from town A at a constant speed of 50 miles an hour. Represent the distance between the cars as a function of time. What if the second car is moving east too? What if it starts 1 hour late?

Example 2.3.6: graphs as curves

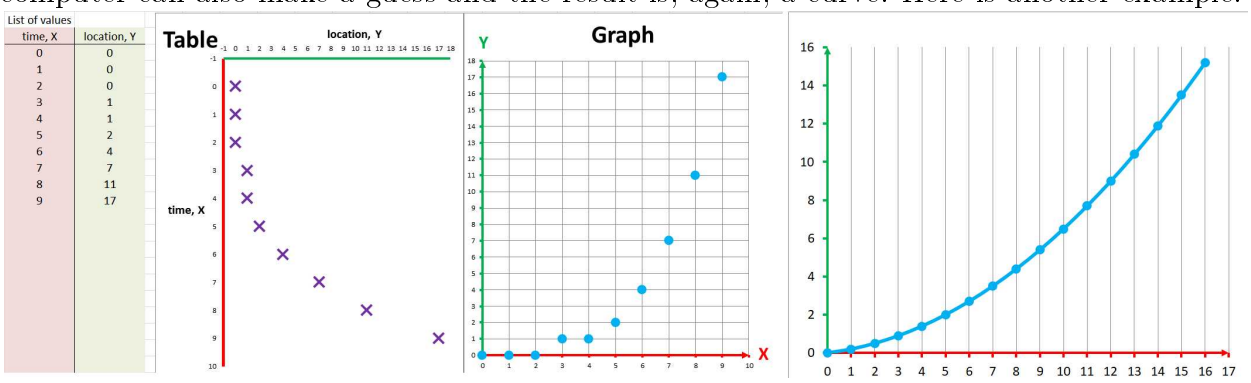
Instead of producing more data, it is common to fill in the gaps in a graph with a stroke of a pen:



... or with a click:



The computer can also make a guess and the result is, again, a curve! Here is another example:



Exercise 2.3.7

Represent a round trip.

By choosing an appropriate set Y of outputs, we can model "motion" through quantities other than locations: temperature, pressure, population, money, etc.

Both functions and sequences can be partially or fully represented by lists of values. In addition, they can also be defined by *formulas*. For example, we can match this sequence and that function:

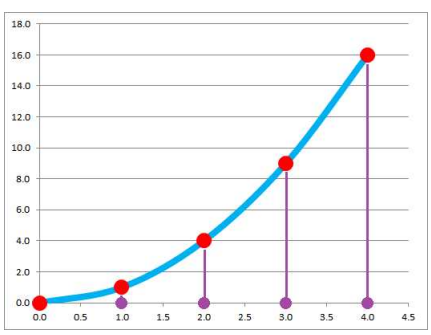
$$a_n = n^2 \quad \text{and} \quad f(x) = x^2 \,.$$

Both are described by the command: "Square the input!" Their tables of values are identical, initially:

n	$y = n^2$
0	0
1	1
2	4
3	9
⋮	⋮

x	$y = x^2$
0	0
1	1
2	4
3	9
⋮	⋮

Unlike the table of the sequence, the table of the function misses more rows, not just at the end but also these: for $x = .5$, $x = \sqrt{2}$, etc. One can also see the difference if we plot the two graphs together:



Between any two input of the sequence (red), the function might have a whole interval of extra inputs (blue).

Thus, every function $y = f(x)$ creates a sequence $a_n = f(n)$, but not necessarily vice versa. A "counter-example" is provided by $a_n = (-1)^n$.

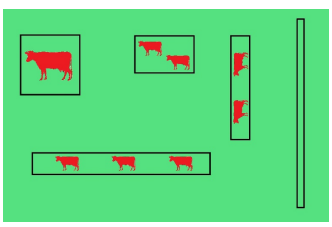
Exercise 2.3.8

Give examples of other sequences that don't produce functions in this manner.

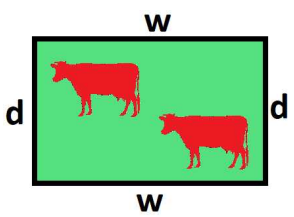
2.4. How numerical functions emerge: optimization

▶ **PROBLEM:** *A farmer with* 100 *yards of fencing material wants to build as large a rectangular enclosure as possible for his cattle.*

The scope of possibilities is infinite:



We are supposed to find the dimensions of this rectangle, the width and the depth:



It makes sense if by "the largest enclosure" we will mean the one with the largest *area*.

We will go through multiple stages, initially relying only on our common sense (and some middle school math).

(1) Trial and error. We start by randomly choosing possible measurements of the enclosure and compute its area with the formula:

$$\text{Area} \;=\; \text{width} \;\cdot\; \text{depth}.$$

We start with a rectangle with width 20 and increase the depth:

- 20 by 20 gives us an area of 400 square yards.
- 20 by 30 gives us an area of 600 square yards.
- 20 by 40 gives us an area of 800 square yards...

Of course, the area is getting bigger and bigger; however, 30 by 30 gives us 900! The pattern is unclear.

We will need to collect more data. Let's speed up this process with a spreadsheet. We introduce the variables:

- w is for the width, and
- d is for the depth, then
- a is for the area:

$$a = w \cdot d\,.$$

(2) Collecting data in a spreadsheet. We first need to compile all possible combinations of the width, column W, and the depth, column D. We choose to go every 10 yards. Then the two quantities, independently, run through these 11 numbers:

$$\text{width} \;= 0,\; 10,\; 20,\; ...,\, 100 \;\; \text{and depth} \;= 0,\; 10,\; 20,\; ...,\, 100\,.$$

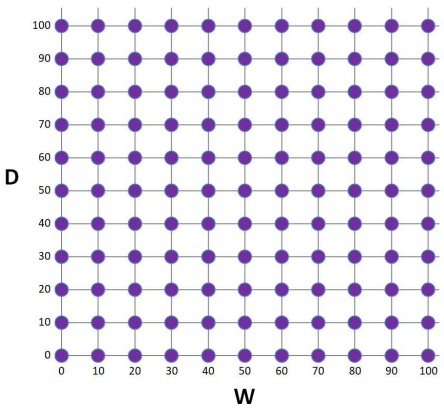
Together, there are $11 \cdot 11 = 121$ possible combinations. The first challenge is to list all possible pairs of width and depths in a spreadsheet. The simplest approach is to fix one value of w, starting with 0, and then

start varying the value of d until we reach 10, then we set w equal to 10 and so on. Once we have them all, we also have all the areas too; we just compute the area, column A, with the formula:

```
=RC[-2]*RC[-1]
```

This is the result (left):

W	D	A
0	0	0
10	0	0
20	0	0
30	0	0
40	0	0
50	0	0
60	0	0
70	0	0
80	0	0
90	0	0
0	10	0
10	10	100
20	10	200
30	10	300
40	10	400
50	10	500
60	10	600
70	10	700
80	10	800
90	10	900
0	20	0



Unfortunately, we can't just look through this column and find the largest number! The reason is that we need to test whether a given combination of width and depth uses exactly 100 yards of the fencing material. Is there a better way? To investigate, let's *plot* these pairs (right). Together, they form an 11×11 *square* of possible combinations, with its width and depth corresponding to the width and depth of the enclosure! It appears that it is better to arrange these pairs in a *table* than in a list.

(3) Establishing sets for the variables. We choose to consider the dimensions every 10 yards via these two sets named after these two quantities:

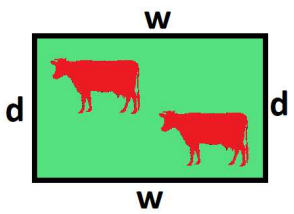
$$W = \{0,\ 10,\ ..., 100\} \quad \text{and} \quad D = \{0,\ 10,\ ..., 100\}\,.$$

The table, and the spreadsheet, takes the form:

$W \backslash D$	0	10	...	100
0	?			
1				
⋮				
100				?

What do we fill it with? We can, and will, fill it with the areas but there is something more urgent: the perimeter of the enclosure. Indeed, such enclosures as 0×0 and 100×100 simply don't make sense!

(4) Establishing a relation between the variables. For simplicity, we assume that we are to use the whole 100 yards of fencing. This makes up its perimeter:



Therefore, a relation between the sets W and D is defined by the following:

► Two numbers w and d are related when they form an rectangle the perimeter p of which is 100:

$$p = 2(w + d) = 100\,.$$

We fill the table with the corresponding values of the perimeter:

```
=2*(RC2+R2C)
```

This is the result:

Good

W\D	0	10	20	30	40	50	60	70	80	90	100
0	0	20	40	60	80	100	120	140	160	180	200
10	20	40	60	80	100	120	140	160	180	200	220
20	40	60	80	100	120	140	160	180	200	220	240
30	60	80	100	120	140	160	180	200	220	240	260
40	80	100	120	140	160	180	200	220	240	260	280
50	100	120	140	160	180	200	220	240	260	280	300
60	120	140	160	180	200	220	240	260	280	300	320
70	140	160	180	200	220	240	260	280	300	320	340
80	160	180	200	220	240	260	280	300	320	340	360
90	180	200	220	240	260	280	300	320	340	360	380
100	200	220	240	260	280	300	320	340	360	380	400

Bad

Next, we mark the acceptable values, equal to 100, of the perimeter in yellow; these are the only ones allowed.

(5) Visualizing the relation. For more accurate results, we need more data; we go every single yard:

$$W = \{0,\ 1,\ 2,\ ...,100\} \quad \text{and} \quad D = \{0,\ 1,\ 2,\ ...,100\}\,.$$

Then the manual data analysis above isn't possible anymore: We have a 101×101 table with $10,201$ pairs. We fill the table with the values of the perimeter p using the same spreadsheet. We then highlight – automatically – the cells where the value is exactly 100:



These are the only ones allowed and they seem to form a straight line!

(6) Computing the quantity to be maximized. We next use the spreadsheet to compute the area of the enclosure with these dimensions according to the formula:

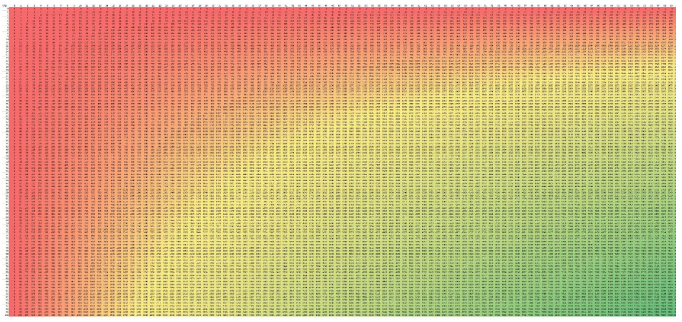
```
=RC2*R2C
```

referring to the same row and second column and the second row and the same columns, respectively. As a result, the table is filled with these values (shown for the 11×11 table):

Areas

W\D	0	10	20	30	40	50	60	70	80	90	100
0	0	0	0	0	0	0	0	0	0	0	0
10	0	100	200	300	400	500	600	700	800	900	1000
20	0	200	400	600	800	1000	1200	1400	1600	1800	2000
30	0	300	600	900	1200	1500	1800	2100	2400	2700	3000
40	0	400	800	1200	1600	2000	2400	2800	3200	3600	4000
50	0	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
60	0	600	1200	1800	2400	3000	3600	4200	4800	5400	6000
70	0	700	1400	2100	2800	3500	4200	4900	5600	6300	7000
80	0	800	1600	2400	3200	4000	4800	5600	6400	7200	8000
90	0	900	1800	2700	3600	4500	5400	6300	7200	8100	9000
100	0	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000

If we bring the yellow line of the allowable perimeters to this table, we notice that the largest allowable areas seem to be between 20×30 and 30×20. Now we consider the bigger table (101×101) and color – automatically – the cells according to the value of the area it contains:



The values grow – from red to green – in the diagonal direction!

(7) Estimating the maximum. Matching the pictures of the perimeters and the areas, we discover that the largest area must be somewhere (halfway?) between the two extremes, 0×50 and 50×0, at the two ends of the yellow line. Maybe even somewhere (halfway?) between 20 and 30. Could it be 25×25?

It's a fair guess but there must be a better way! The problem is that selecting the allowable data from the whole table of pairs of w and d is too cumbersome. It would help if we had a direct flow of data from w (or d) to a, as a function. To that end, we represent the relation between W and D as a *function*.

(8) Establishing a function from the relation. We express d in terms of w. We take our relation $2(w + d) = 100$ and solve it for d:

$$d = 50 - w\,.$$

Then there is exactly one d for each w; it's a function! Here is its list of values (10 yards at a time):

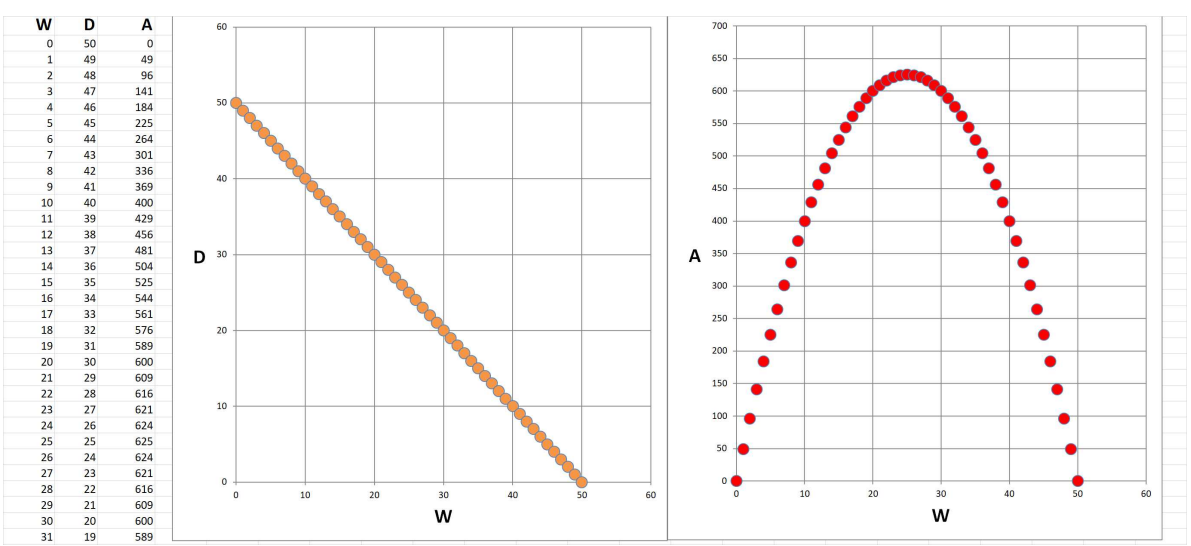
w	d
0	50
10	40
20	30
30	20
40	10
50	0

If we choose to go 1 yard at time, we have 51 rows and we have to put those in a new spreadsheet. The first column is for the width w running through: 0, 1, 2, ..., 50. The second is for the depth d, evaluated by

```
=50-RC[-1]
```

referring to the previous column.

(9) Visualizing the function. These two columns are seen on the left below and the graph of the depth as a function of the width is a straight line (middle):



W	D	A
0	50	0
1	49	49
2	48	96
3	47	141
4	46	184
5	45	225
6	44	264
7	43	301
8	42	336
9	41	369
10	40	400
11	39	429
12	38	456
13	37	481
14	36	504
15	35	525
16	34	544
17	33	561
18	32	576
19	31	589
20	30	600
21	29	609
22	28	616
23	27	621
24	26	624
25	25	625
26	24	624
27	23	621
28	22	616
29	21	609
30	20	600
31	19	589

The area a is also computed in the next column as the product of w and d:

```
=RC[-2]*RC[-1]
```

and then plotted against the width (right). Since there is exactly one a for each w, this is also a function! Its graph is a curve.

(10) Estimating the maximum. Looking at the last plot, $w = 25$ seems to be a clear choice with the corresponding area $a = 25 \cdot 25 = 625$ square yards. This confirms our previous observations.

Exercise 2.4.1

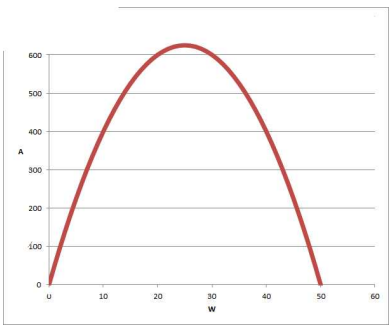
What happens if, instead, we express w in terms of d?

The problem seems to be solved, but unfortunately, the plot has *gaps*! What if we have overlooked a width that it gives us the area bigger than 625?

Let's consider the function that we have built in this spreadsheet: a depends on w only. What is this function? With more middle school algebra, we make this function explicit:

$$a = wd = w(50 - w)\,.$$

We can now easily plot 100 or 100,000 points at as small increments as we like:



With no visible gaps, the answer remains the same:

$$w = 25,\ a = 625\,.$$

Are there invisible gaps? How can we be sure? The graph of a quadratic function, such as

$$a = w(50 - w) = -w^2 + 50w\,,$$

is a *parabola* (Chapter 4). Its "tip", called the vertex, lies half-way between these two points (0 and 50):

$$x = \frac{0 + 50}{2} = 25\,.$$

Let's review our solution.

We named the quantities that appear in the initial problem and then translated its sentences into algebra. The result was the following *optimization problem*:

> ▶ Find such values of w and d that $0 \le w \le 50$, $0 \le d \le 50$ and $a = wd$ is the largest, subject to the relation $w + d = 50$.

Then using the function $d = 50 - w$ derived from the relation to eliminate d from the problem by substitution:

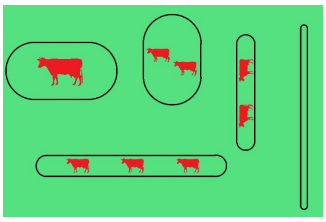
> ▶ Find such a value of w that $0 \le w \le 50$ and $a = w(50 - w)$ is the largest.

Exercise 2.4.2

Solve a modified problem: A river is adjacent to the enclosure, which will have, consequently, *three* sides.

Exercise 2.4.3

Solve a modified problem with a new kind of enclosures required by the problem: Semicircles are attached to the rectangles:



We've solved the problem, but our knowledge is too limited when functions are more complex than just a quadratic polynomial. Calculus (Chapter 2DC-3) will help.

Example 2.4.4: optimization

Find two numbers whose difference is 100 and the product is a minimum.

Step 1. Deconstruct:
1. two numbers, whose
2. difference is 100, and
3. the product is a minimum.

Translate:
1. introduce the "variables": x is the first number, y is the second number;
2. constraint: $x - y = 100$;
3. p is their product: $p = xy$, minimize p.

This is a math problem now.

Step 2. Eliminate the extra variables to create a function of single variable to be maximized or minimized. The constraint, an equation connecting the variables, is

$$x - y = 100\,.$$

Solve the equation for y:

$$y = x - 100\,,$$

and eliminate y from p by substitution:

$$p = xy = x(x - 100)\,.$$

Step 3. Optimize this function:

$$p(x) = x(x - 100) = x^2 - 100x\,.$$

The two end-points of the interval are 0 and 100; therefore, the vertex of this parabola corresponds to:

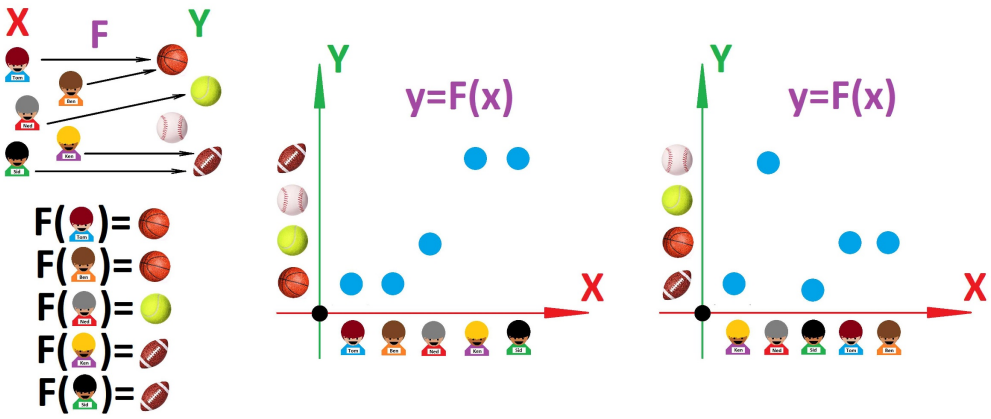
$$x = 50\,.$$

Step 4. Provide the answer using the original language of the problem: Substitute x into y, as follows:

$$\begin{aligned} y &= x - 100 \\ &= 50 - 100 \\ &= -50\,. \end{aligned}$$

Answer: The two numbers are 50 and -50.

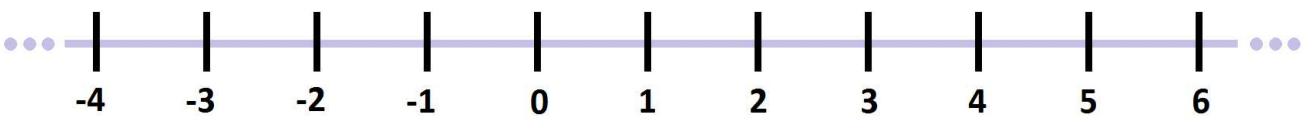
This is the summary of the analysis of the function that represents the preferences of the boys for different games presented earlier in the chapter:



The two graphs represent the same function! They only look different because we have rearranged the elements of the domain, X, and the codomain, Y. Such a move is no longer possible when we deal with *numerical functions* because numbers have an inherent structure, an order.

We use the tables and the graphs of functions to discover patterns in the data. However, this is only possible when the sets themselves have *structure*. For example, a deck of cards remains the same deck after it's been shuffled but there is also a hierarchical relation within the deck that makes all the difference to the players.

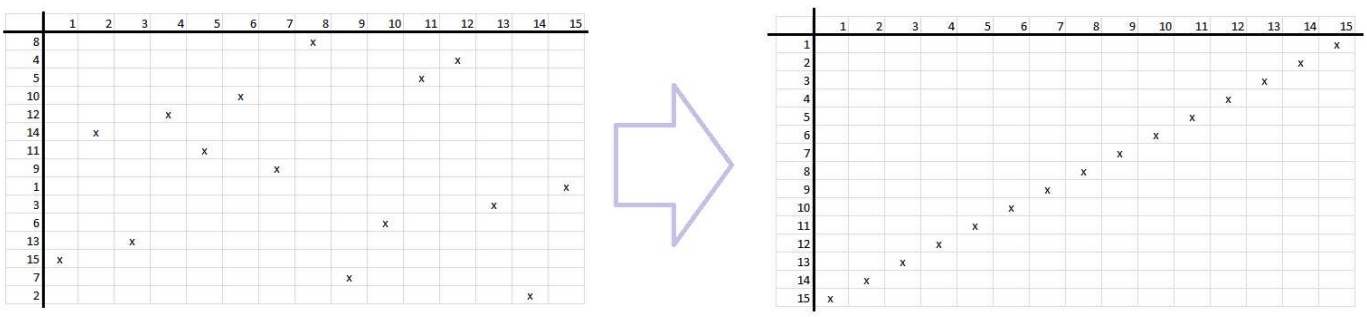
The simplest example of a set with a structure is a set of *locations* on a straight road. We choose milestones to be such as a set. It is their order that makes it impossible to reshuffle them without losing important information. We will use that to our advantage. We visualize the set of milestones as markings on a straight line, according to their *order* ($... < 1 < 2 < 3 < ...$):



The same representation is also used for *time*. Every marking on a line (another line!) indicates a moment of time when some repeatable event, such as a bell ringing or a clock's hand passing a particular position, occurs.

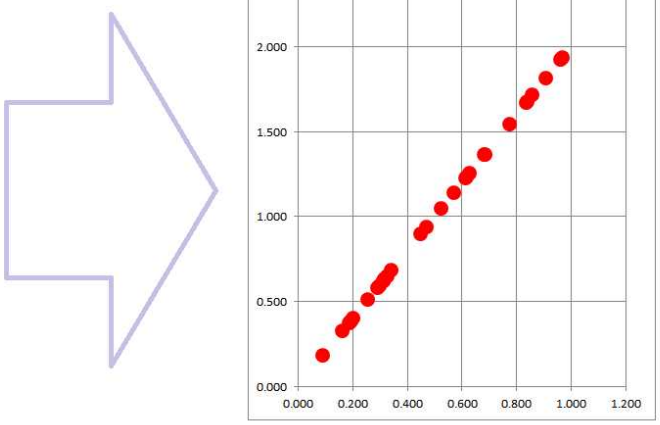
Example 2.4.5: hidden patterns

The table of the function on the left has no apparent pattern – until we re-arrange the rows according to this order:



Similarly, a seemingly random list of pairs of numbers, x and $y = F(x)$, produces a straight line when plotted against properly ordered numbers:

X	Y
0.255	0.510
0.970	1.939
0.617	1.234
0.189	0.378
0.908	1.817
0.325	0.650
0.291	0.583
0.315	0.630
0.311	0.622
0.163	0.326
0.296	0.592
0.614	1.229
0.091	0.181
0.188	0.376
0.859	1.718
0.966	1.932
0.448	0.895
0.193	0.386
0.835	1.671
0.683	1.366
0.684	1.368
0.628	1.255
0.311	0.622
0.524	1.048
0.838	1.675
0.569	1.139



2.5. Set building

So far, all numerical sets in this chapter have been *subsets of the set of real numbers* **R**. In particular, numerical sets emerge as domains and codomains of numerical functions. They may also come from *solving equations*.

Example 2.5.1: sets from equations

Unless entirely nonsensical, every statement in mathematics is true or false:

$$\begin{aligned} 1+1 &= 2 \quad \texttt{TRUE} \\ 1+1 &= 3 \quad \texttt{FALSE} \end{aligned}$$

What about this:

$$x+1 = 2 \quad \texttt{TRUE OR FALSE?}$$

It depends on x, of course. We can, therefore, use equations to form sets.

Consider these:

- We face the equation $x+2=5$. After some work, we find: $x=3$.
- We face the equation $3x=15$. After some work, we find: $x=5$. Is there more?
- We face the equation $x^2-3x+2=0$. After some work, we find: $x=1$. Is that it?
- We face the equation $x^2+1=0$. After all the work, we can't find any x. Should we keep trying?

Here, x is a *label* that stands for an unspecified number that is meant to satisfy this condition. In other words, we seek such numbers that, when they replace x in the equation, we see a true statement. It could simply be trial and error.

Let's take the first equation:

$$x+2=5\,.$$

We replace x with a number above or write a number in the blank square below:

$$\square+1=5\,.$$

For example:

- Is $x=1$ a solution? Plug it in the equation: $x+2=5$ becomes $(1)+2=5$. `FALSE`. This is *not* a solution.

- Is $x = 2$ a solution? Plug it in the equation: $x + 2 = 5$ becomes $(2) + 2 = 5$. `FALSE`. This is *not* a solution.
- Is $x = 3$ a solution? Plug it in the equation: $x + 2 = 5$ becomes $(3) + 2 = 5$. `TRUE`. This *is* a solution.
- Should we stop now? Why would we? For all we know, there may be more solutions.

We never say that we have found "the" solution unless we know for sure that there is only one.

Exercise 2.5.2

Interpret each of these equations as a relation.

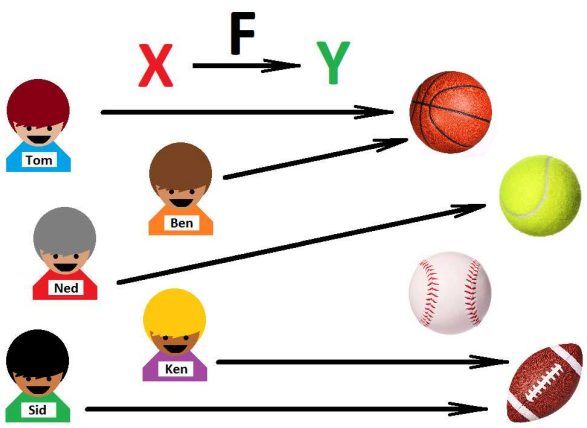
Exercise 2.5.3

Solve these equations:

$$(a)\ x^2 + 2x + 1 = 0, \quad (b)\ \frac{x}{x} = 0, \quad (c)\ \frac{x}{x} = 1\,.$$

But what does it *mean* to solve an equation? We have tried to find x that satisfies the equation... But what are we supposed to have at the end of our work?

Let's go back to our running example of boys and balls:



It tells us what game each boy prefers. What about the other way around? Which boys prefer a particular game?

- Which boys prefer basketball? The answer isn't "Tom", and it isn't "Ben"; it's "Tom and Ben and nobody else".
- Which boys prefer tennis? "just Ned".
- Which boys prefer baseball? "No one".
- Which boys prefer football? "Ken and Sid only".

But each question – one for each element of the codomain Y – is also an *equation*:

- Find x with $F(x) =$ basketball. Tom is a solution, and Ben is a solution. Combined, Tom and Ben are *the* solutions.
- Find x with $F(x) =$ tennis. Ned is the solution.
- Find x with $F(x) =$ baseball. No solutions.
- Find x with $F(x) =$ football. Ken and Sid are the solutions.

This is how we understand this idea:

▶ *A solution of an equation with respect to x is an element that, when put in the place of x in the equation, gives us a true statement.*

However, we must present *all* x's that satisfy the equation. In other words, the answer is a *set*:

- The solution set of the equation $F(x) =$ basketball is { Tom, Ben }.
- The solution set of the equation $F(x) =$ tennis is { Ned }.
- The solution set of the equation $F(x) =$ baseball has non elements.
- The solution set of the equation $F(x) =$ football is { Ken, Sid }.

All of these sets are subsets of the domain X. This is the terminology we will routinely use.

Definition 2.5.4: equation and its solution

Suppose $f : X \to Y$ is a function and b is one of the elements of Y. The *solution set* of the equation $f(x) = b$ is the set of all x in X that make the equation true. To *solve an equation* means to find its solution set.

In other words, the solution set is *the* solution!

Next, the sets above can be presented in this spirit:

Set-building notation

$$\{\, x : \quad \text{condition for } x \,\}$$

The expression stands for the set of *all* x that satisfy the condition. What kind of condition? An equation, as above. Any condition as long as it is specific enough for us to unambiguously answer the question "does x satisfy it?". For example:

$$\{ \text{ student: 20 years old } \}.$$

The set from which we pick x's one at a time is presented or assumed to be known.

Warning!

Many sources also use:

$$\{x| \quad \text{condition for } x \,\}.$$

For example, the equations above are seen as conditions. Below we list their solution sets (left):

$$\begin{aligned}
\{x,\ \text{boy} \ :\ F(x) =\ \text{basketball}\ \} &= \{\ \text{Tom, Ben}\ \} \\
\{x,\ \text{boy} \ :\ F(x) =\ \text{tennis}\ \} &= \{\ \text{Ned}\ \} \\
\{x,\ \text{boy} \ :\ F(x) =\ \text{baseball}\ \} &= \{\ \ \} \\
\{x,\ \text{boy} \ :\ F(x) =\ \text{football}\ \} &= \{\ \text{Ken, Sid}\ \}
\end{aligned}$$

These description can sometimes be *simplified* (right). One can imagine that we simply went over the list of X and tested each of its elements. The third one is special; it has no elements. The following notation will be routinely applied to this set.

Empty set

$$\varnothing$$

Exercise 2.5.5

Show that the empty set is a subset of any set.

Exercise 2.5.6

Simplify the following sets:

$$\begin{array}{ll} \{x,\ \text{boy} & :\ \text{his shirt is red }\} \\ \{y,\ \text{ball} & :\ \text{is preferred by two boys }\} \\ \{y,\ \text{ball} & :\ \text{is round }\} \end{array}$$

Example 2.5.7: inclusion vs. implication

Here is an interpretation of the definition of subset. The definition says that $A \subset B$ when the following is satisfied:

▶ `IF` x belongs to A, `THEN` x belongs to B;

or

▶ x belongs to $A \implies x$ belongs to B.

The converse of this implication is false when $A \neq B$. Furthermore, we have the following:

▶ x satisfies the condition for $A \implies x$ satisfies the condition for B.

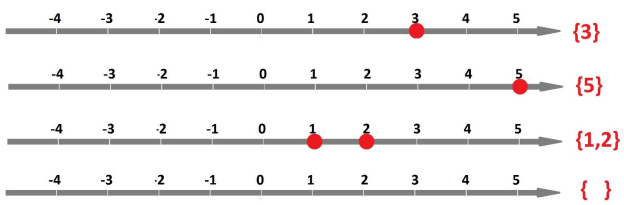
In other words, the latter condition is less restrictive.

Example 2.5.8: solution sets of equations

Let's take another look at the equations above, assuming that the "ambient" set is the set of real numbers:

equation:	answer?	solution set:
$x+2=5$	$x=3$	$\{3\}$
$3x=15$	$x=5$	$\{5\}$
$x^2-3x+2=0$	$x=1$ and...	$\{1,2\}$
$x^2+2x+1=0$	no x?	$\{\ \ \}$

This is how we visualize these four sets:



Below we use the set-building notation again on the left, and then on the right, we see another, simpler, representation of the set:

$$\begin{array}{ll} \{x:\ x+2=5\} & =\{3\} \\ \{x:\ 3x=15\} & =\{5\} \\ \{x:\ x^2-3x+2=0\} & =\{1,2\} \\ \{x:\ x^2+1=0\} & =\{\ \ \}=\varnothing \end{array}$$

The simplest way to represent a set is, of course, a list.

Exercise 2.5.9

Solve these equations:

$$x=x, \quad 1=1, \quad 1=0\,.$$

Example 2.5.10: sets from inequalities

Unless entirely nonsensical, every statement in mathematics is true or false:

$$\begin{array}{ll} 1\leq 2 & \texttt{TRUE} \\ 1\leq 0 & \texttt{FALSE} \end{array}$$

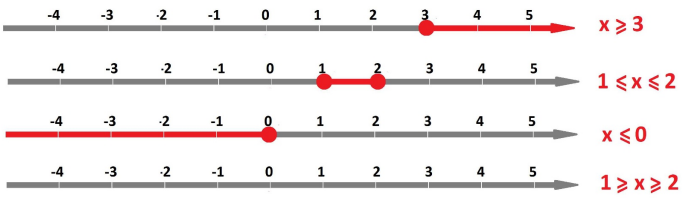
What about this:

$$1\leq x \quad \texttt{TRUE OR FALSE?}$$

It depends on x, of course. We can, therefore, use inequalities to form sets:

$$\begin{array}{llll} \{x, & \text{real}: & 3 \leq & x & \} \\ \{x, & \text{real}: & 1 \leq & x < 2 & \} \\ \{x, & \text{real}: & & x \leq 0 & \} \\ \{x, & \text{real}: & 1 \geqslant & x \geqslant 2 & \} \end{array}$$

These are also subsets of the real number line:



Because it's infinite, as we zoom in on the real line, we see just as many numbers as before. This is the reason why there is no such thing as the list of all real numbers! Since we can't test them one by one, *visualization* becomes especially important.

Exercise 2.5.11

Solve these inequalities:

$$x \leq x, \quad x < x, \quad 1 \leq 1, \quad 1 < 0\,.$$

When inequalities are involved, the set-building notation is used along with a more compact notation. The following will be universally used.

Definition 2.5.12: interval

A non-empty subset of the real line defined by an inequality or two inequalities is called an *interval*.

We start with sets of real numbers located between two specified numbers $a < b$:

Interval notation, segments

$\{x: \ a \leq x \leq b\}$	$=$	$[a,b]$	*closed bounded interval*
$\{x: \ a \leq x < b\}$	$=$	$[a,b)$	*half-open or half-closed bounded interval*
$\{x: \ a < x \leq b\}$	$=$	$(a,b]$	*half-open or half-closed bounded interval*
$\{x: \ a < x < b\}$	$=$	(a,b)	*open bounded interval*

Warning!

Sometimes "finite interval" is used instead.

The terminology is in the right column:

- The bracket – "[" or "]" – is used in the interval notation when the number adjacent to it is included in the set. It uses a *non-strict* inequality $\leq$.
- The parenthesis – "(" or ")" – is used when the number is excluded from the set. It uses a *strict* inequality $<$.

Second, one of the ends may be infinite. In other words, we consider all numbers on one side of a real number a or b:

Interval notation, rays

$\{x:$	$a \le x < \infty$	$\}$	$=$	$[a, +\infty)$	*closed unbounded interval*
$\{x:$	$a < x < \infty$	$\}$	$=$	$(a, +\infty)$	*open unbounded interval*
$\{x:$	$-\infty < x \le b$	$\}$	$=$	$(-\infty, b]$	*closed unbounded interval*
$\{x:$	$-\infty < x < b$	$\}$	$=$	$(-\infty, b)$	*open unbounded interval*

Warning!

Sometimes "infinite interval" is used instead.

Third, both of the ends may be infinite. In other words, we consider the real line:

Interval notation, line

$$\{x: \quad -\infty < x < +\infty \quad \} \quad = \quad (-\infty, +\infty) \quad = \mathbf{R}$$

Infinity is always excluded from any interval because...

Warning!

Infinity is not a number.

Thus, an interval is a subset of the real line that consists of all the numbers between two numbers or between a number and infinity.

Since x is assumed to be a real number, some inequalities that involve infinities, such as $-\infty < x$, are redundant.

Example 2.5.13: lists of numbers

If we limit ourselves to the integers here, the same inequalities won't produce intervals but *lists*, some of them infinite, for example:

$\{x,$ integer:	$1 \le x \le 4$	$\}$	$=$	$\{1,2,3,4\}$	$\{x,$ integer:	$1 \le x$	$\}$	$=$	$\{1,2,3,4,...\}$
$\{x,$ integer:	$1 \le x < 4$	$\}$	$=$	$\{1,2,3\}$	$\{x,$ integer:	$1 < x$	$\}$	$=$	$\{2,3,4,...\}$
$\{x,$ integer:	$1 < x \le 4$	$\}$	$=$	$\{2,3,4\}$	$\{x,$ integer:	$x \le 4$	$\}$	$=$	$\{...,1,2,3,4\}$
$\{x,$ integer:	$1 < x < 4$	$\}$	$=$	$\{2,3\}$	$\{x,$ integer:	$x < 4$	$\}$	$=$	$\{...,1,2,3\}$

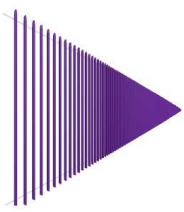
Exercise 2.5.14

Simplify:

$$\{x > 0: \quad \text{is an integer } \}.$$

Example 2.5.15: point of infinity?

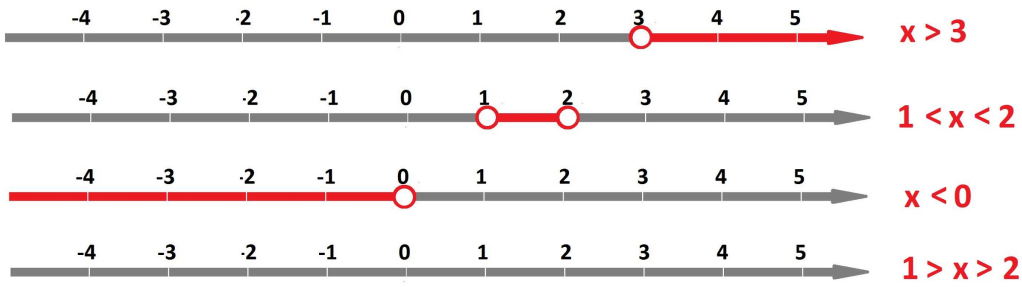
Infinity is not a number nor a location. It is, therefore, never included in the sets of numbers that we consider. One can *perceive* infinity as, for example, a "point" where a long fence disappears or where two railroad tracks meet on the horizon:





We may attempt to "approach" infinity while staying within the set of numbers (Chapter 2DC-1).

For visualization, we use a little circle to indicate a missing point at the end of an interval:



In the last section, we saw an optimization problem that required finding the *largest* possible output of a function. But the outputs form a set! We will be often concerned with finding largest and smallest elements of sets of numbers:



The following definition puts optimization problems in a proper context.

Definition 2.5.16: minimum and maximum of set

Suppose X is a set of real numbers.

- The *minimum* of X is the smallest element of X; i.e., it is such a number m in X that
$$m \leq x \quad \texttt{FOR EACH } x \text{ in } X\,.$$
- The *maximum* of X is the largest element of X; i.e., it is such a number M in X that
$$x \leq M \quad \texttt{FOR EACH } x \text{ in } X\,.$$

They are denoted, respectively, by:

$$\min X \quad \text{and} \quad \max X$$

Warning!

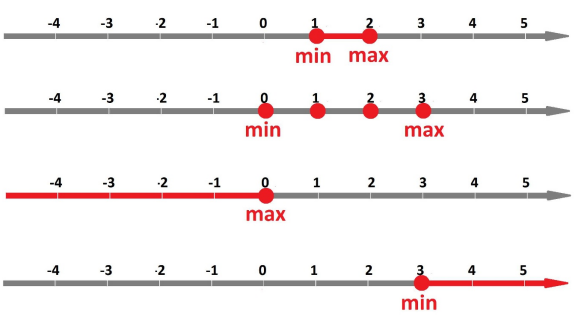
Both minimum and maximum must be elements of the set.

Example 2.5.17: max and min

Here's a simple example:

$$\min[a,b] = a\,, \quad \max[a,b] = b\,.$$

Here's another:



However, $(a,b]$ has no minimum and $[a.b)$ has no maximum.

Exercise 2.5.18

Explain the grammar in the definition – why "the maximum"?

Example 2.5.19: max-min of lists

An even simpler case is a list of numbers arranged in increasing order; then the task is easy:

$$\min\{-1,3,7,12,16\} = -1\,, \quad \max\{-1,3,7,12,16\} = 16\,.$$

We just examine the list and pick the largest or the smallest element. However, if the list grows infinite and *unbounded*, such as:

$$X = \{1,2,3,4,5,...\}\,,$$

there is, once again, no maximum! We can't say that the maximum is the infinity because...

Exercise 2.5.20

Finish the sentence.

A set may have no maximum or no minimum! However, even when there isn't the best one, there may be many candidates.

Definition 2.5.21: bounds of set

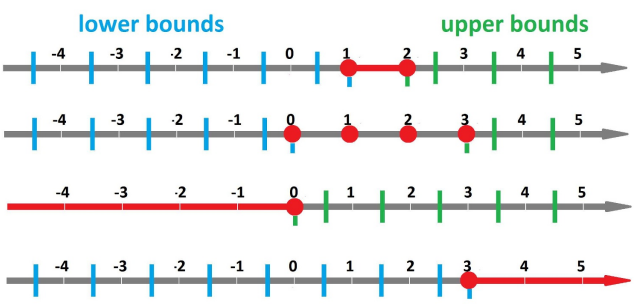
Suppose X is a set of real numbers.

- A number b (it doesn't have to belong to X) is called a *lower bound* of X if
$$b \leq x \quad \texttt{FOR EACH } x \text{ in } X\,.$$
- A number B (it doesn't have to belong to X) is called an *upper bound* of X if
$$x \leq B \quad \texttt{FOR EACH } x \text{ in } X\,.$$

A set that possesses both lower and upper bounds is called *bounded*, otherwise *unbounded*.

Example 2.5.22: bounds of sets

There are infinitely many bounds unless the set is infinite on that side:



Example 2.5.23: boundedness of sets

These are bounded sets:

$$\{-1,3,7,12,16\},\ [a,b],\ (a,b)\,,$$

and these are unbounded:

$$\mathbf{R},\ \mathbf{Z},\ \{1,2,3,4,5,...\}\,.$$

Exercise 2.5.24

Show that if a set X is bounded, it is a subset of an interval $[A, B]$, where A and B are any of its lower and upper bounds, respectively.

Exercise 2.5.25

What are the max, min, and bounds of the empty set $\varnothing$? What about $\mathbf{R}$?

Exercise 2.5.26

State the converse of "the maximum is an upper bound" and find out if it is true.

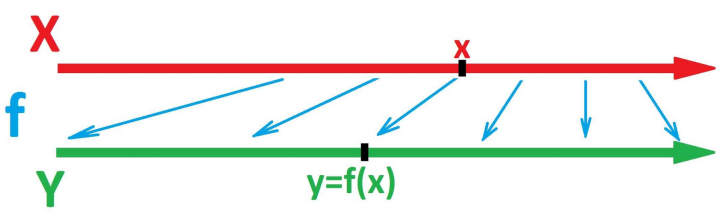
Exercise 2.5.27

Solve the following equations:

$$(a)\ x^2+2x+1=0\,,\quad (b)\ x^2=-1\,,\quad (c)\ x^2=1\,,\quad (d)\ \frac{x}{x}=1\,,\quad (e)\ \frac{x}{x}=0\,.$$

2.6. The xy-plane: where graphs live...

A relation, or a function, deals with two sets of numbers: the domain X and the codomain Y. That's why we need two axes, one for $X=\mathbf{R}$ and one for $Y=\mathbf{R}$. How do we arrange them? We can use the method presented above: putting the exes next to each other and connecting them by arrows:



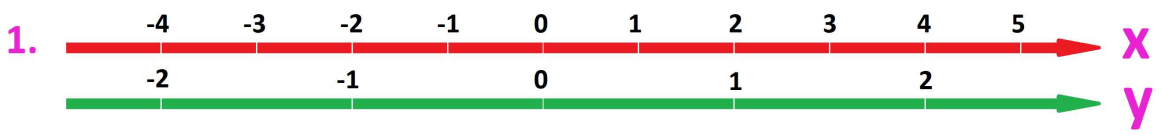
But since $X=\mathbf{R}$ is infinite, however, we would need infinitely many arrows. Is there a better way? We already know another approach: *tables and graphs*. Instead of side-by-side, we place X horizontally and Y vertically.

Step 1

We start with a *real line* $\mathbf{R}$, or the x-axis (Chapter 1). That's where the real numbers live, and now X and Y are subsets of $\mathbf{R}$. So, we will need two *copies* of the real line. We give them special names:

- the *x-axis*, and
- the *y-axis*.

Just as the inputs and the outputs of a function have typically nothing to do with each other (such as time vs. space, or space vs. temperature), the two axes may have different unit segments:

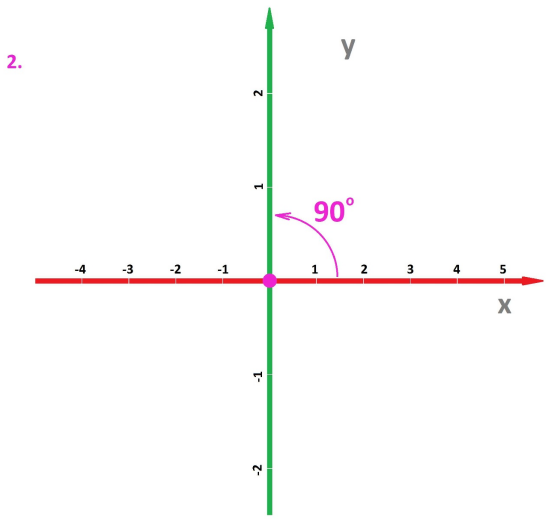


Step 2

To move toward the table we need, we arrange the two coordinate axes in a typical way as follows:

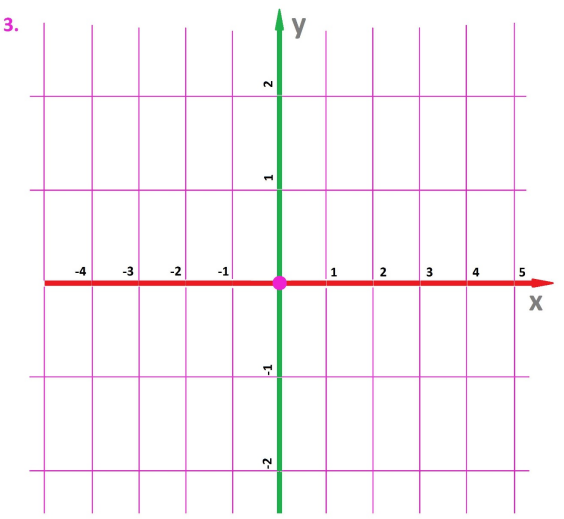
- The x-axis is horizontal, with the positive direction pointing right.
- The y-axis is vertical, with the positive direction pointing up.

Usually, the two axes are attached to each other at their origins:



Step 3

Finally, we attach "fabric" to this "frame". We use the marks on the axes to draw a *rectangular grid*.

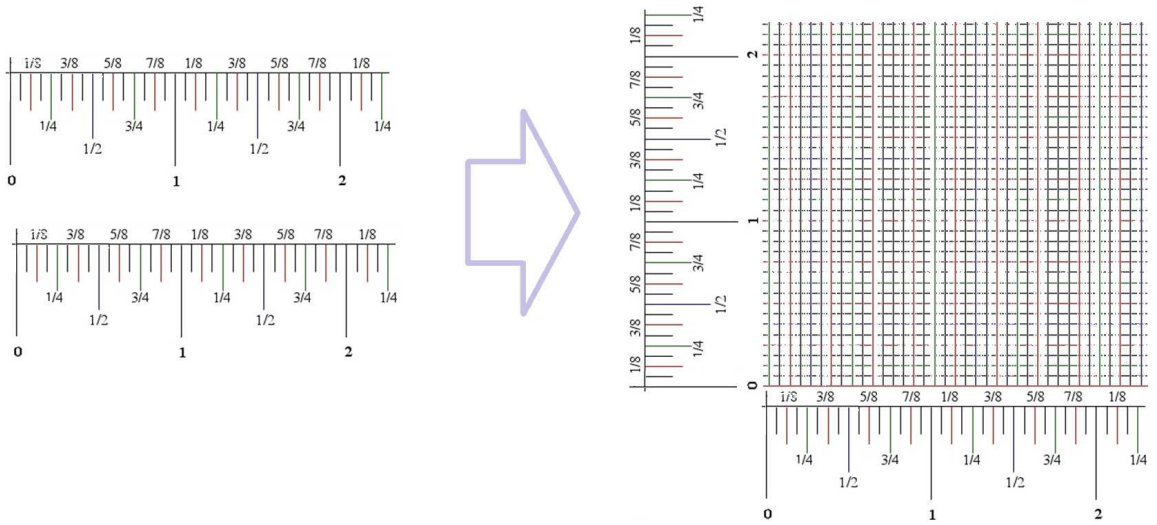


Now we have what we call the *Cartesian plane*, or simply the *xy-plane*. As it is made from a combination of two copies of $\mathbf{R}$ and is often denoted as follows:

xy-plane
$\mathbf{R}^2$

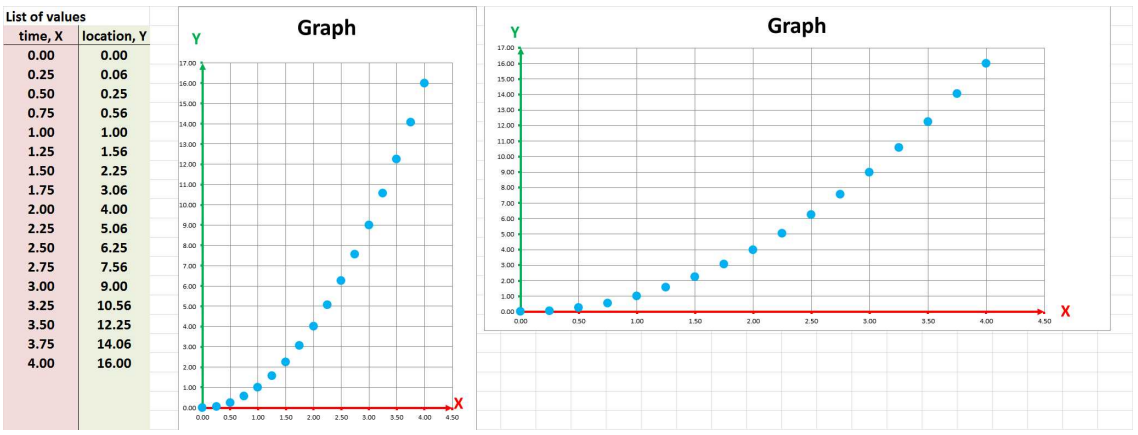
The notation is also explained by the fact that the area of an $r \times r$ square is $r \cdot r = r^2$. This is literally true, however, only when both axes measure length.

The idea that the real line is like a ruler leads to the idea that the xy-plane is like a *ruled paper*:



Example 2.6.1: resizing graphs

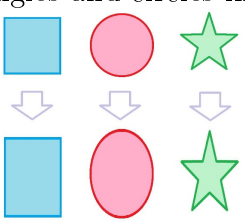
In the context of plotting graphs, it is frequently the case that the relative dimensions of x and y are unimportant, and then the xy-plane can be resized arbitrarily and *disproportionately*. The graphs change too! The chart in this spreadsheet shows how different the graph of the same function might look:



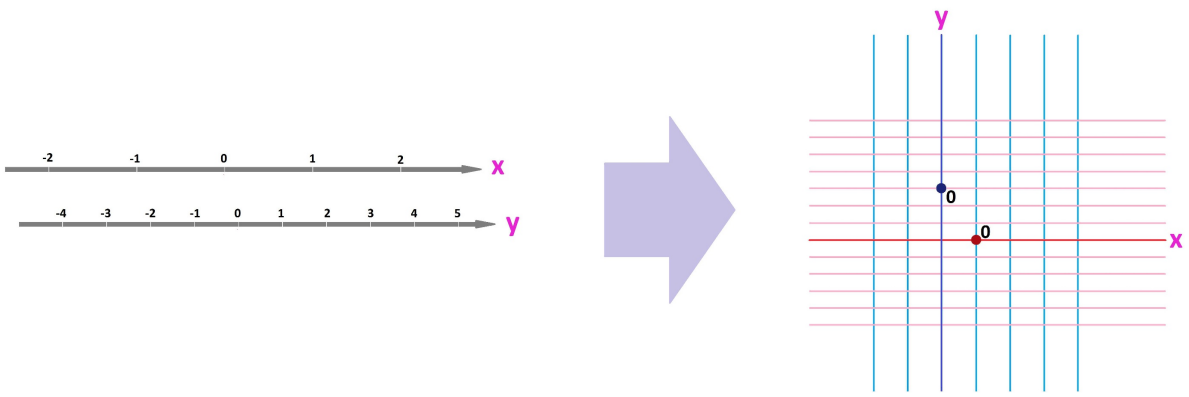
List of values

time, X	location, Y
0.00	0.00
0.25	0.06
0.50	0.25
0.75	0.56
1.00	1.00
1.25	1.56
1.50	2.25
1.75	3.06
2.00	4.00
2.25	5.06
2.50	6.25
2.75	7.56
3.00	9.00
3.25	10.56
3.50	12.25
3.75	14.06
4.00	16.00

Such resizing will turn squares into rectangles and circles into ovals:



This fact imposes an important limit on how well the graph visualizes the function. The size, of course, doesn't matter. The angles might be telling us nothing; for example, the inclination – up or down – of the graph can't disappear under this re-sizing but its steepness can change. We can't, therefore, say that this line is "steep" but only that it is "steeper" than another one plotted on the same coordinate plane. In this context, it is also often acceptable to have the origins of the two axes misaligned or even absent:

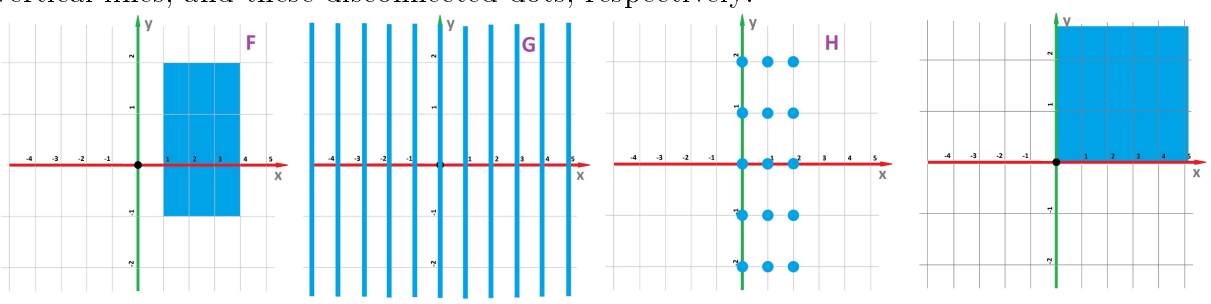


Example 2.6.2: locations of graphs

Let's see what we can we say about the locations of the graphs of functions with these domains and codomains:

$$F : [1,4] \to [-1,2], \quad G : \mathbf{Z} \to \mathbf{R}, \quad H : \{0,1,2\} \to \mathbf{Z}, \quad Q : [0,\infty) \to [0,\infty)\,.$$

Even though we have no other information, we can point out that the graphs lie within this rectangle, these vertical lines, and these disconnected dots, respectively:



The idea of the *Cartesian coordinate system* is similar to the one for the real line, but this time there are *two* axes and *two* coordinates for each point. We use the above setup to produce a correspondence:

location P $\longleftrightarrow$ a pair of numbers (x, y)

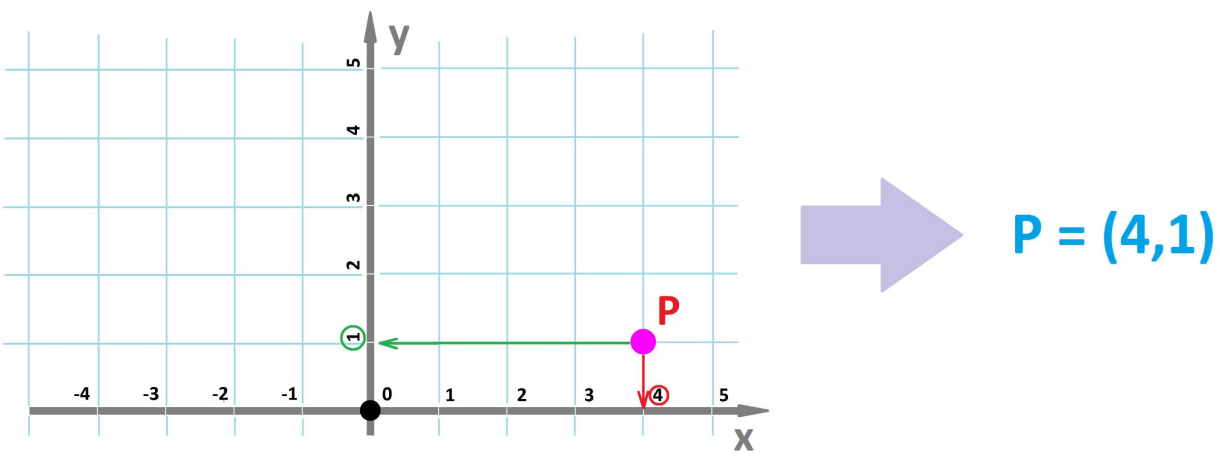
It works in *both directions*:

1. $\to$ First, suppose P is a *location* on the plane. We then draw a vertical line through P until it intersects the x-axis. The mark, x, of the location where they cross is the *x-coordinate* of P. We next draw a horizontal line through P until it intersects the y-axis. The mark, y, of the location where they cross is the *y-coordinate* of P.

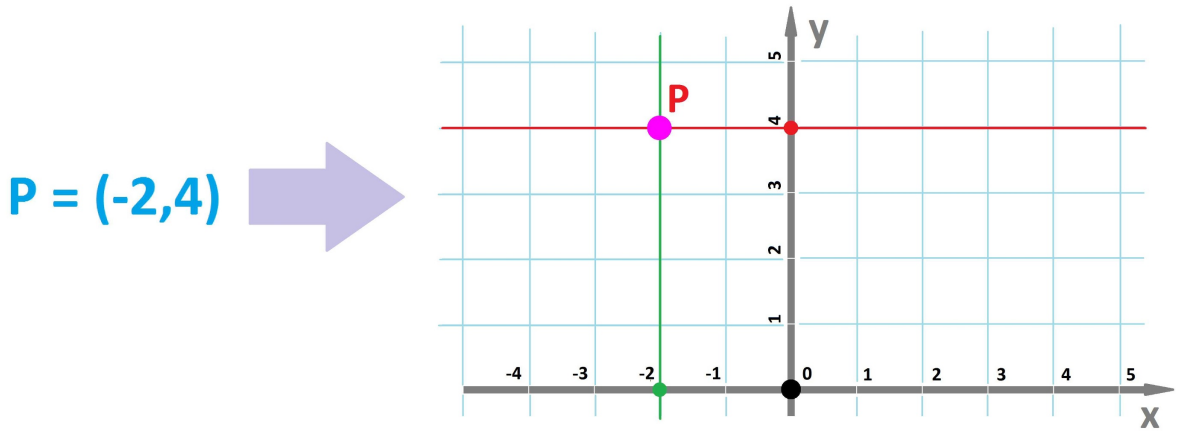
2. $\leftarrow$ On the flip side, suppose x and y are *numbers*. First, we find the mark x on the x-axis and draw a vertical line through this point. Second, we find the mark y on the y-axis and draw a horizontal line through this point. The intersection of these two lines is the corresponding *location* P on the plane.

Example 2.6.3: coordinates

We illustrate this idea below with a specific example. From a point to its coordinate:



From coordinates to a point:



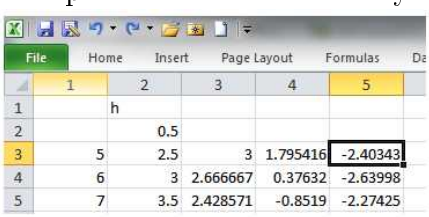
The notation is as follows:

xy-coordinates

$$\left(x - \text{coordinate}\,,\ y - \text{coordinate} \right)$$

Example 2.6.4: coordinates in computing

The 2-dimensional Cartesian system isn't as widespread as the 1-dimensional (numbers!). It is, however, common in certain areas of computing. Spreadsheet applications use the Cartesian system – starting at the upper right corner – to provide a convenient way of representing locations of cells:



This is the difference:

	spreadsheet	Cartesian system
1st coordinate	rows `R`, down	x, right
2nd coordinate	columns `C`, right	y, up

This idea is instrumental in helping us express a value located in one cell in terms of values located in other cells by reference. For example, this formula computes the double of the number contained in the cell located at the intersection of row 3 and column 5:

```
=2*R3C5
```

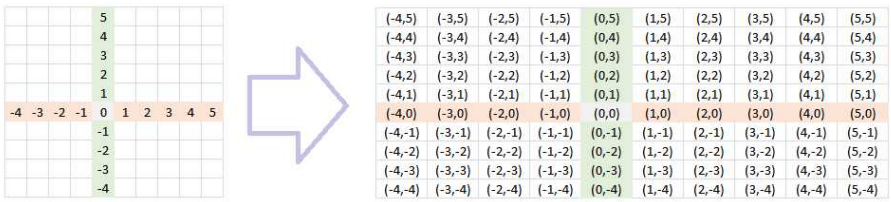
This formula computes the sum of the number contained in the cell in row 5 and column 2 and the number contained in the cell in row 1 and column 5:

```
=R5C2+R1C5
```

There are also relative references. For example, this formula takes the number contained in the cell located 3 rows up and 2 columns right from the current cell:

```
=R[-3]C[2]
```

Note that the row number, which is y, comes first. In contrast, this is what the proper Cartesian system for spreadsheets would look like:



Thus, every point on the xy-plane is, or can be, labeled with a pair of numbers.

Warning!

The notation (a, b) is, unfortunately, the same for the point on the xy-plane with coordinates a and b and for the open interval from a to b.

Once the coordinate system is in place, it is acceptable to think of locations as pairs of numbers, and vice versa. In fact, we can write:

$$P = (x, y)\,.$$

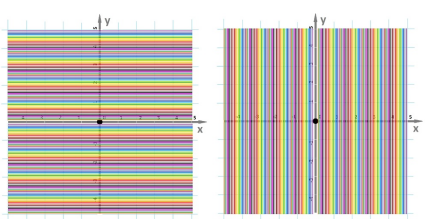
It is important to realize that what we are dealing with is a *set* too! This is the set of all pairs of real numbers presented in the set-building notation:

$$\mathbf{R}^2 = \{(x, y) :\ x \text{ real},\ y \text{ real}\}\,.$$

The xy-plane is just a visualization of this set. Below we consider some of its simplest *subsets*.

Example 2.6.5: lines are fibers of plane

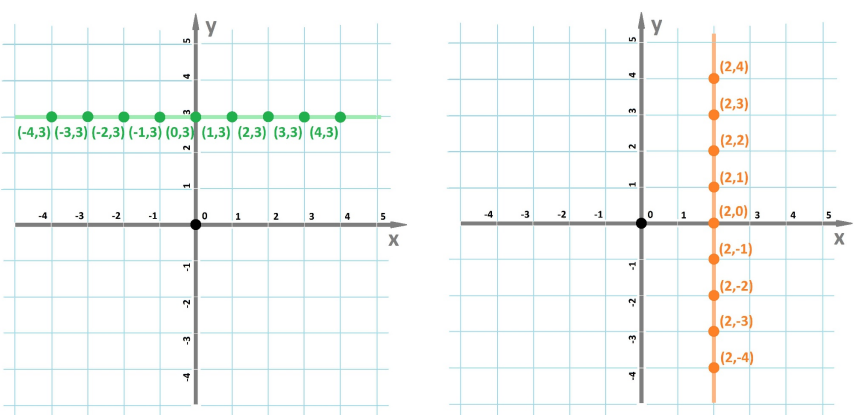
One can think of the xy-plane as a *stack* of lines, vertical or horizontal, each of which is just a shifted copy of the corresponding axis:



We can use this idea to reveal the internal structure of the coordinate plane.

- If L is a line parallel to the x-axis, then all points on L have the same y-coordinate. Conversely, if a set L of points on the xy-plane consists of all points with the same y-coordinate, L is a line parallel to the x-axis.
- If L is a line parallel to the y-axis, then all points on L have the same x-coordinate. Conversely, if a set L of points on the xy-plane consists of all points with the same x-coordinate, L is a line parallel to the y-axis.

Two specific examples are shown below:

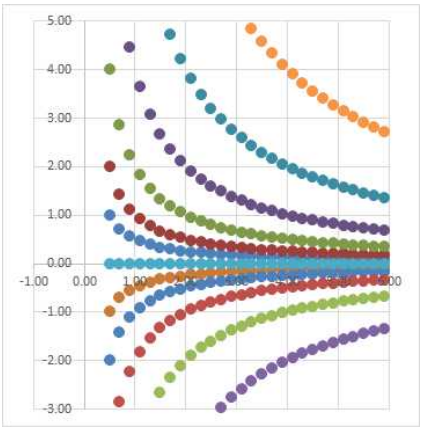


Then, we have a compact way to represent these two lines:

- horizontal: $y = 3$, and
- vertical: $x = 2$.

Such an equation removes a degree of freedom! On the line, there is only one, and we are left with a single point. With two degrees of freedom on the plane, we have a line.

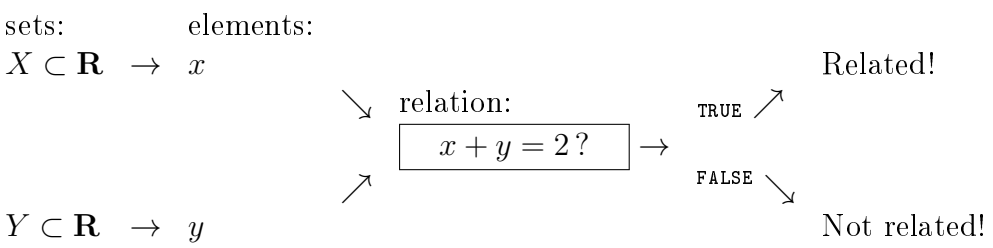
The most important subsets of the plane will be initially the graphs of relations and functions. For example, sequences of numbers produce sequences of points on the plane:



How to use this setup to do *geometry* is explained in Chapter 5.

2.7. Linear relations

Recall that a relation between two sets is any pairing of their elements. This time, the sets are sets of numbers and the condition to be checked is an equation:



So, a numerical relation processes a pair of numbers x and y and tells only one thing: related or not related. For example:

$$\begin{array}{llll} x=1,\ y=2 & \to & 1+2=2? \ \texttt{FALSE} & \to \text{ Not related!} \\ x=1,\ y=1 & \to & 1+1=2? \ \texttt{TRUE} & \to \text{ Related!} \end{array}$$

Warning!

Just because both sets might be the sets of real numbers ($X = \mathbf{R}$ and $Y = \mathbf{R}$), we shouldn't think of the relation as one of a set with *itself*.

Example 2.7.1: maximizing enclosure

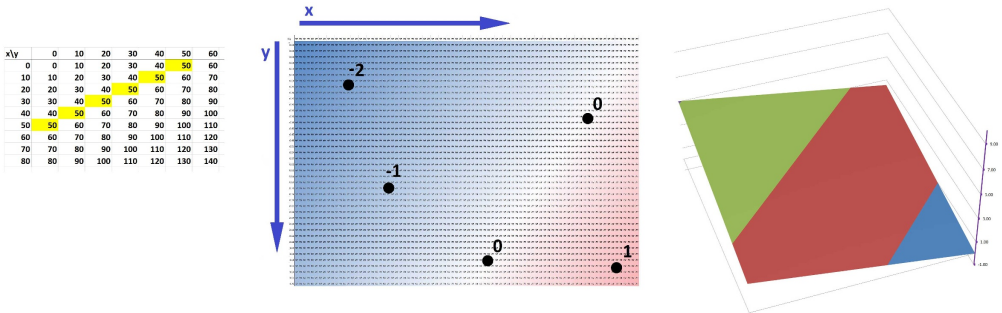
This is how we visualized a relation earlier in this chapter. The cattle enclosure has width x and height y and the two are related – because of the fixed amount of material for the fence available – by a relation:

$$x + y = 50\,.$$

To speed up the analysis, we pre-computed *all* values of $x + y$ for every eligible pair x and y. The result is a table filled by means of the following spreadsheet formula:

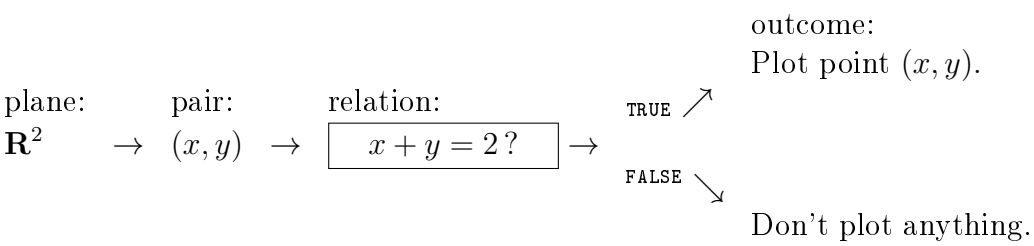
```
=RC2+R2C
```

It was easy in a small table to color the cells with value of $x + y$ equal to 50 (left); they form a line:



We can furthermore color a very large array of cells (middle): the color of the (x, y)-cell is determined by the value of $x + y$. The linear pattern still seems conceivable. The value of $x + y$ can also be visualized as the elevation at that location (right).

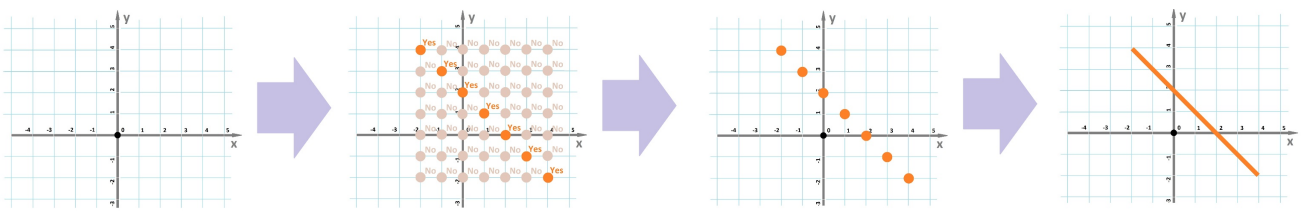
What is the scope of possible inputs in the above diagram? Any value of x is possible and, independently, any value of y. Therefore, all *pairs* (x, y) are possible. Furthermore, *plotting the graph* of a numerical relation means processing a pair of numbers (x, y), one at a time, and producing an output, which is: related or not related, Yes or No, `TRUE` or `FALSE`, a point or no point. For example:



We can try to do this by hand, one at a time:

$$(0,0) \to \text{ No! } \quad (1,0) \to \text{ No! } \quad (1,1) \to \text{ Yes!...}$$

It takes a lot of tries to produce a picture that reveals a pattern:



On the far right, we show our conjecture about the graph of the relation; it looks like a straight line!

Exercise 2.7.2

Show that the equation $2x + 2y = 4$ represents the same relation!

In the last section, we saw two examples of relations the graphs of which are lines:

- The relation $y = c$ produces a horizontal line because every point (x, y) is plotted as long as $y = c$ (there is no restriction on x).
- The relation $x = a$ produces a vertical line because every point (x, y) is plotted as long as $x = a$ (there is no restriction on y).

As a summary, we give a precise definition:

Definition 2.7.3: graph of relation

Suppose R is a relation between two sets X and Y of real numbers. Then, the *graph* of R is the set of all points on the xy-plane with x and y related by R:

$$\text{graph of } R \;=\; \{(x, y) :\; x \;\text{ is related to }\; y\}\,.$$

We use the set-building notation! This relation is, typically, an equation, and in this case, "most" of the points on the plane won't satisfy it. Those that do will likely form a *curve*.

We start with the simplest, and the most common, kind.

Theorem 2.7.4: Graph of Linear Relation

The graph of any linear relation, i.e.,

$$Ax + By = C\,,$$

with either A or B not equal to zero, is a straight line.

Exercise 2.7.5

What is the graph when $A = B = 0$? Hint: There are two cases.

Exercise 2.7.6

State the converse of the theorem and find out if it's true.

It is called an *implicit equation of the line*. When we represent the line by a function (next section), the equation becomes *explicit*.

The ideas of *linear algebra* have a very humble beginning.

Example 2.7.7: linear equation

Suppose we have a type of coffee that costs \$3 per pound. How much do we get for \$60?

The setup is the following. Let x be the weight of the coffee. Since the total price is 60, we have a *linear equation*:

$$3x = 60\,.$$

We solve it:

$$x = \frac{60}{3} = 20\,.$$

The operations are very simple, and the complexity comes from elsewhere: the number of variables.

Example 2.7.8: mixtures

Suppose we have the Kenyan coffee that costs \$2 per pound and the Colombian coffee that costs \$3 per pound. How much of each do you need to have 6 pounds of blend with the total price of \$14?

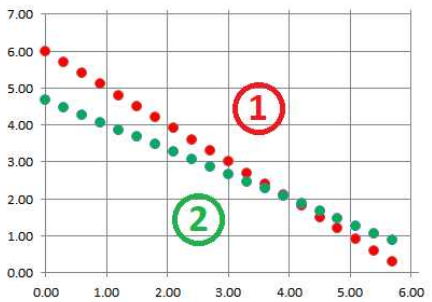
We don't try to solve the problem. We "translate" it into mathematics. Let x be the weight of the Kenyan coffee, and let y be the weight of Colombian coffee. Since the total weight is 6, we have a linear relation between x and y:

$$\boxed{1} \quad x + y = 6\,.$$

Since the total price of the blend is \$14, we have another linear relation between x and y:

$$\boxed{2} \quad 2x + 3y = 14\,.$$

According to the theorem, the graphs of the relations are lines, these lines:



Then, for a combination of weights x and y to satisfy *both* of the requirements, the point (x, y) has to belong to *both* of the lines! The solution, therefore, is the point that they have in common, called their *intersection* (Chapter 3). This point can be guessed to be $(x, y) = (4, 2)$. This can be confirmed by substituting the two numbers $x = 4,\ y = 2$ into the two relations:

$$\begin{array}{|c|} \hline 1 \\ \hline 2 \\ \hline \end{array} \quad \begin{array}{llllllll} x & +y & = 6 & \to & 4 & +2 & = 6 & \texttt{TRUE} \\ 2x & +3y & = 14 & \to & 2 \cdot 2 & +3 \cdot 2 & = 14 & \texttt{TRUE} \end{array}$$

An algebraic solution may be as follows. From the first equation, we derive: $y = 6 - x$ (it's a function!). Then substitute this y into the second equation: $2x + 3(6 - x) = 14$. Solve this new equation: $-x = -4$, or $x = 4$. Substitute this back into the first equation: $(4) + y = 6$, then $y = 2$.

Such a problem is called a *system of linear equations* (further considered in Chapter 4DE-2):

$$\begin{cases} x & +y & = 6\,, \\ 2x & +3y & = 14\,. \end{cases}$$

We can collect the data in tables as follows:

$$\begin{array}{|ll|l|} \hline 1 \cdot x & +1 \cdot y & = 6 \\ \hline 2 \cdot x & +3 \cdot y & = 14 \\ \hline \end{array}\ , \text{ rewritten: } \begin{array}{|c|c|c|c|c|c|c|c|c|} \hline 1 & \cdot & x & + & 1 & \cdot & y & = & 6 \\ \hline 2 & \cdot & x & + & 3 & \cdot & y & = & 14 \\ \hline \end{array}\ , \text{ rewritten: } \left[\begin{array}{cc|c} 1 & 1 & 6 \\ 2 & 3 & 14 \end{array}\right].$$

The 2-by-2 (left) part of the resulting table is made of the coefficients of x and y in the equations. It is called a *matrix*.

Exercise 2.7.9

Set up a system of linear equations – but do not solve it – for the following problem: "An investment portfolio worth $\$1,000,000$ is to be formed from the shares of: Microsoft - \$5 per share, and Apple - \$7 per share. If you need to have twice as many shares of Microsoft than Apple, what are the numbers?"

Exercise 2.7.10

Set up, do not solve, the system of linear equations for the following problem: "One serving of tomato soup contains 100 cal and 18 g of carbohydrates. One slice of whole bread contains 70 cal and 13 g of carbohydrates. How many servings of each should be required to obtain 230 cal and 42 g of carbohydrates?"

Exercise 2.7.11

Solve the system of linear equations:

$$\begin{cases} x - y & = 2\,, \\ x + 2y & = 1\,. \end{cases}$$

Exercise 2.7.12

Solve the system of linear equations and geometrically represent its solution:

$$\begin{cases} x - 2y &= 1\,, \\ x + 2y &= -1\,. \end{cases}$$

Exercise 2.7.13

Geometrically represent this system of linear equations:

$$\begin{cases} x - 2y &= 1\,, \\ x + 2y &= -1\,. \end{cases}$$

Exercise 2.7.14

What if there is a third type of coffee in the example, say \$4 per pound?

2.8. Relations vs. functions

There may be many relations with the same graph:

$$\begin{array}{ccccc} & & x+y-2=0 & & \\ & \nwarrow & \uparrow & \nearrow & \\ x+x+y+y=4 & \leftarrow & \boxed{x+y=2} & \rightarrow & y=-x+2 \\ & \swarrow & \downarrow & \searrow & \\ -y=x-2 & & -x=y-2 & & x=-y+2 \end{array}$$

But two of them are special: they are *functions*. One is y in terms of x and the other x in terms of y.

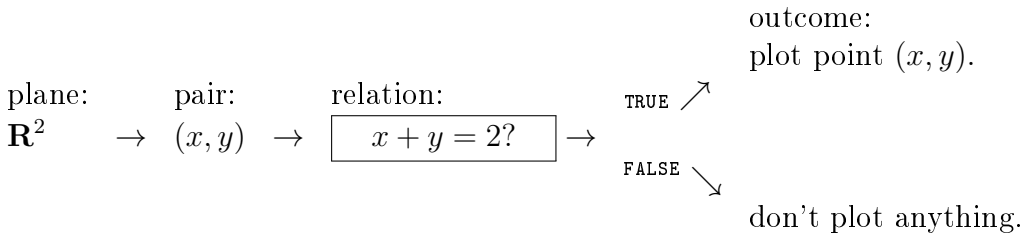
It is crucial that *graphs are sets* too; they are subsets of $\mathbf{R}^2$. In fact, we can still use the set-building notation:

$$\{(x,y):\ \text{condition on } x,\ y\}\,.$$

This condition, just as before, is often an equation; for example:

$$\{(x,y):\ x^2+y^2=1\}\,.$$

Because of the indirect nature of the definition of this set, plotting the graph of a numerical relation is cumbersome:



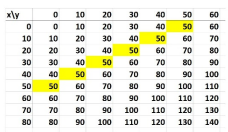
What else can we do?

Example 2.8.1: linear relation

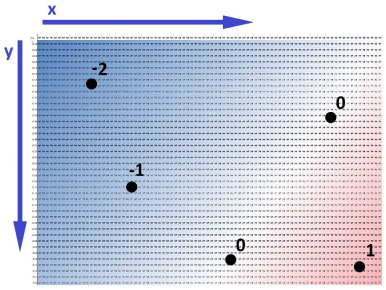
This is how we visualized a relation earlier in this chapter. We pre-computed *all* values of $x+y$:

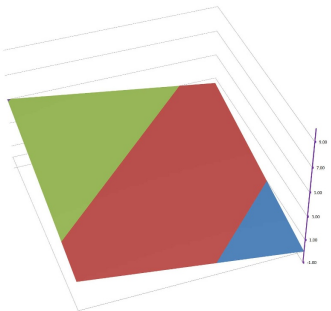
▶ Each value of $x+y$ is placed at the location (x,y).

We filled a spreadsheet with these values and then we examined the patterns: what sets do the cells with constant values form. We discovered them to be *straight lines*:



x\y	0	10	20	30	40	50	60
0	0	10	20	30	40	50	60
10	10	20	30	40	50	60	70
20	20	30	40	50	60	70	80
30	30	40	50	60	70	80	90
40	40	50	60	70	80	90	100
50	50	60	70	80	90	100	110
60	60	70	80	90	100	110	120
70	70	80	90	100	110	120	130
80	80	90	100	110	120	130	140





The yellow line is $x + y = 50$. But it was just a conjecture...

We proved that the graph of any *linear* relation, i.e.,

$$Ax + By = C\,,$$

with either A or B not equal to zero, is a straight line. It is called an *implicit equation of the line*. When we represent the line by a function (below), the equation becomes *explicit*; all we need is to solve for x or for y.

Theorem 2.8.2: When Linear Relation Is Function

A linear relation between the sets $X = \mathbf{R}$ and $Y = \mathbf{R}$,

$$Ax + By = C\,,$$

may be represented by a function, called a linear function, as follows:

1. *When $B \neq 0$, it is a function $F : X \to Y$ given by*

$$y = F(x) = -\frac{A}{B}x + \frac{C}{B}\,.$$

2. *When $A \neq 0$, it is a function $G : Y \to X$ given by*

$$x = G(y) = -\frac{B}{A}y + \frac{C}{A}\,.$$

Indeed, every function is a relation but not every relation is a function, but when it is, there might be two.

Example 2.8.3: two functions in a relation

In the relation $x + y = 50$ above, we have $A = 1$ and $B = 1$, so we can do both:

$$y = -x + 50 \quad \text{and} \quad x = -y + 50\,.$$

In the relation $y = 3$, we have $A = 0$ and $B = 1$; this is the former case and the function is constant:

$$F(x) = 3\,.$$

In the relation $x = 2$, we have $A = 1$ and $B = 0$; this is the latter case and the function is constant:

$$G(y) = 2\,.$$

Exercise 2.8.4

Find all linear functions is these linear relations: (a) $3x - 2y = 2$, (b) $2x = 3$, (c) $-y = 7$.

Exercise 2.8.5

Prove the theorem.

Exercise 2.8.6

What lines are *not* included in case 1? case 2?

Exercise 2.8.7

State both cases of the theorem as implications (an "if-then" statement).

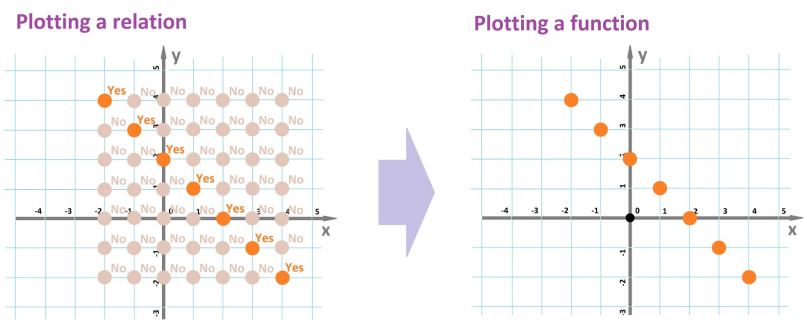
Transitioning to *functions* makes the plotting task much easier. The 49 computations are reduced to just 7:

x	-2	-1	0	1	2	3	4
-2	-4	-3	-2	-1	0	1	2
-1	-3	-2	-1	0	1	2	3
0	-2	-1	0	1	2	3	4
1	-1	0	1	2	3	4	5
2	0	1	2	3	4	5	6
3	1	2	3	4	5	6	7
4	2	3	4	5	6	7	8

$\rightarrow$

x	$y = -x + 2$
-2	4
-1	3
0	2
1	1
2	0
3	-1
4	-2

Instead of testing a lot of points trying, and mostly failing, to find the ones that fit the equation, we just plug in as many values of x as necessary, producing a point every time:



The price we paid was algebra, solving for y:

$$x + y = 2 \implies y = x - 2\,.$$

In general, instead of having to run through a whole *plane* of (x, y)'s – for relations, we only need to run through a *line* of x's – for functions. We also observe that since there can be only *one* point of the graph of a function above each x, the graph of a function must be *one-point thick*, just like a curve.

Exercise 2.8.8

State the converse of "the graph of every linear function is a straight line" and find out if it's true.

Example 2.8.9: circle as relation

Let's consider a more complex relation:

input: (x, y) $\rightarrow$ relation: $x^2 + y^2 = 1?$ $\rightarrow$

TRUE $\nearrow$ output: Plot point (x, y).

FALSE $\searrow$ Don't plot anything.

We quickly run through a few:

$$(0,0) \to \text{No!} \quad (1,0) \to \text{Yes!} \quad (1,1) \to \text{No!} \quad (0,1) \to \text{Yes!} \quad \ldots$$

Let's arrange the initial results in a 2×2 square around the origin:

locations:
$$\begin{array}{|ccc|}\hline (-1,1) & (0,1) & (1,1) \\ (-1,0) & (0,0) & (1,0) \\ (-1,-1) & (0,-1) & (1,-1) \\ \hline\end{array}\,, \quad \text{relation:} \quad \begin{array}{|ccc|}\hline \text{No} & \text{Yes} & \text{No} \\ \text{Yes} & \text{No} & \text{Yes} \\ \text{No} & \text{Yes} & \text{No} \\ \hline\end{array}\,, \quad \text{graph:} \quad \begin{array}{|ccc|}\hline \circ & \bullet & \circ \\ \bullet & \circ & \bullet \\ \circ & \bullet & \circ \\ \hline\end{array}$$

Is this a curve? It seems that we need many more values to get an idea! Unfortunately, when we try to go .5 at a time,

$$x = -1.5,\ -.5,\ .5,\ 1.5 \ \text{ and } \ y = -1.5,\ -.5,\ .5,\ 1.5\,,$$

we just produce No's exclusively. This will continue to happen no matter what how small a step we choose...

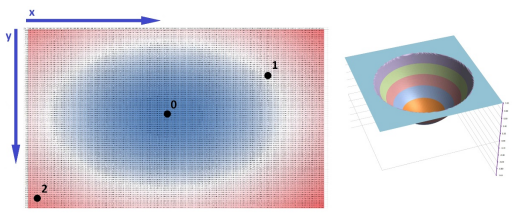
We take another approach to visualization of the graph of a relation. Just as with the linear relations in the last section, we consider *many relations at the same time*:

$$x^2 + y^2 = k\,,$$

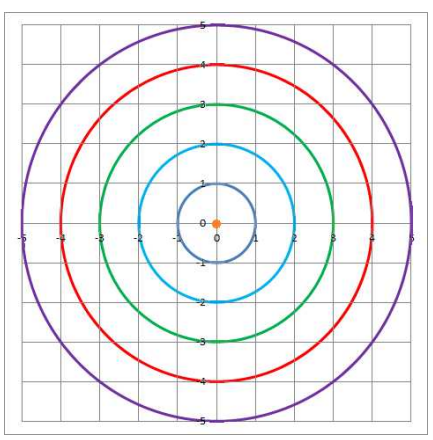
with variable k's. We create a table of the values of the expression on the left in a spreadsheet with the formula:

```
=RC1^2+R1C^2
```

We then color the cells:



The negative values of k are in blue and the positive are in red. Our interest is $k = 1$ and, therefore, the points in blue are too close to the origin and those in red are too far. We choose the white! The circular pattern is clear. The pattern seems to be made from concentric circles with the radius varying with k:



To justify the last conclusion, let's ask: Is there a *function* that represents this relation? Let's solve for y. We have

$$y = \pm\sqrt{1 - x^2}\,.$$

But this is *not* a function; indeed, input $x = 0$ produces *two* outputs: $y = 1$ and $y = -1$. But what does this "equation" represent? This is just a new relation! Indeed, this is the original relation:

▶ x and y are related when

$$x^2 + y^2 = 1\,,$$

and this is the new relation:

▶ x and y are related when

$$y = \pm\sqrt{1 - x^2}\,.$$

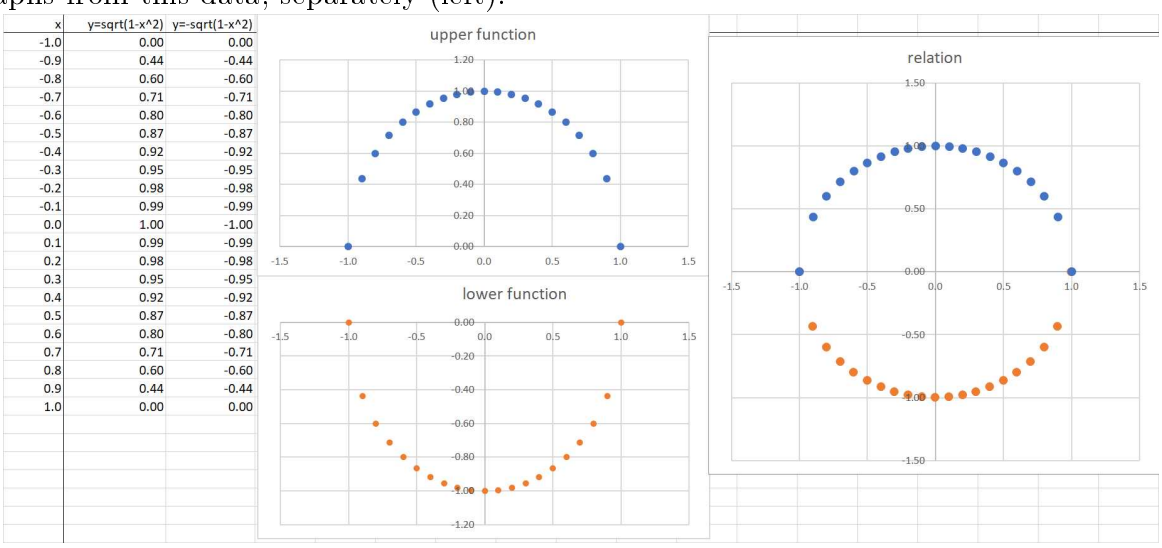
The latter is just another representation of the former.

There is a better way to write it:

▶ x and y are related when

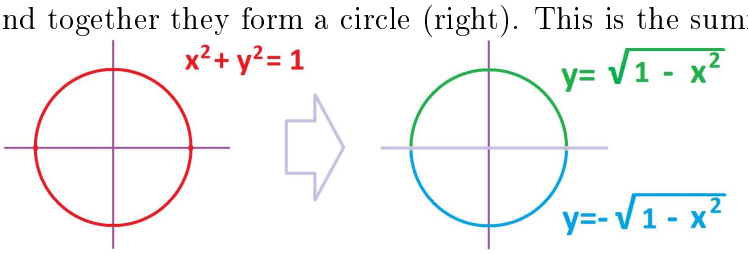
$$y = \sqrt{1 - x^2} \quad \texttt{OR} \quad y = -\sqrt{1 - x^2}\,.$$

How does this help? An examination reveals that we have *two* separate functions. We can then produce a table of values of either one for a large number of x's between -1 and 1. And we can plot their graphs from this data, separately (left):



x	y=sqrt(1-x^2)	y=-sqrt(1-x^2)
-1.0	0.00	0.00
-0.9	0.44	-0.44
-0.8	0.60	-0.60
-0.7	0.71	-0.71
-0.6	0.80	-0.80
-0.5	0.87	-0.87
-0.4	0.92	-0.92
-0.3	0.95	-0.95
-0.2	0.98	-0.98
-0.1	0.99	-0.99
0.0	1.00	-1.00
0.1	0.99	-0.99
0.2	0.98	-0.98
0.3	0.95	-0.95
0.4	0.92	-0.92
0.5	0.87	-0.87
0.6	0.80	-0.80
0.7	0.71	-0.71
0.8	0.60	-0.60
0.9	0.44	-0.44
1.0	0.00	0.00

Either graph is an arc and together they form a circle (right). This is the summary:



Example 2.8.10: hyperbolas

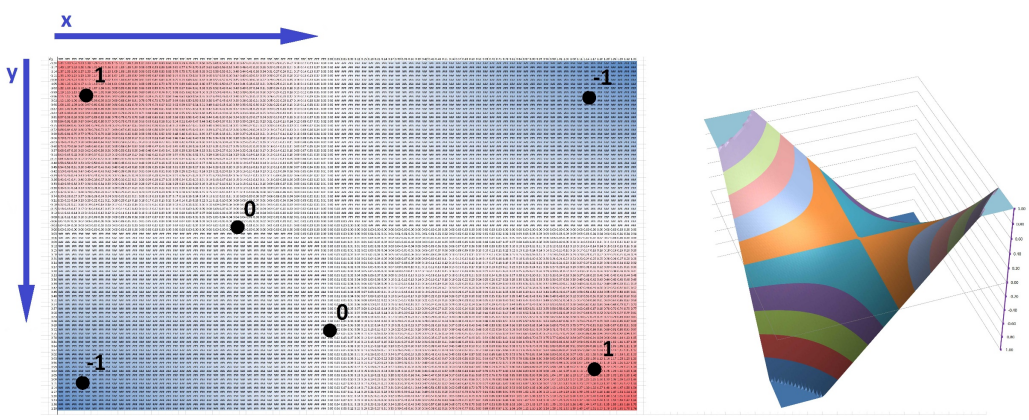
What is the graph of the relation:

$$xy = 1\ ?$$

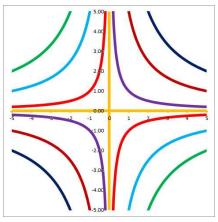
We follow the approach outlined in the last example. We color the cells of an array according to the values of

$$xy = k\,,$$

as follows:



This is what the graphs of these relations look like plotted for various k's; they are curves called *hyperbolas* (Chapter 4):



They are the graphs of the functions that come from solving the equation for y or x:

$$y = \frac{k}{x} \quad \text{and} \quad x = \frac{k}{y}\,.$$

Exercise 2.8.11

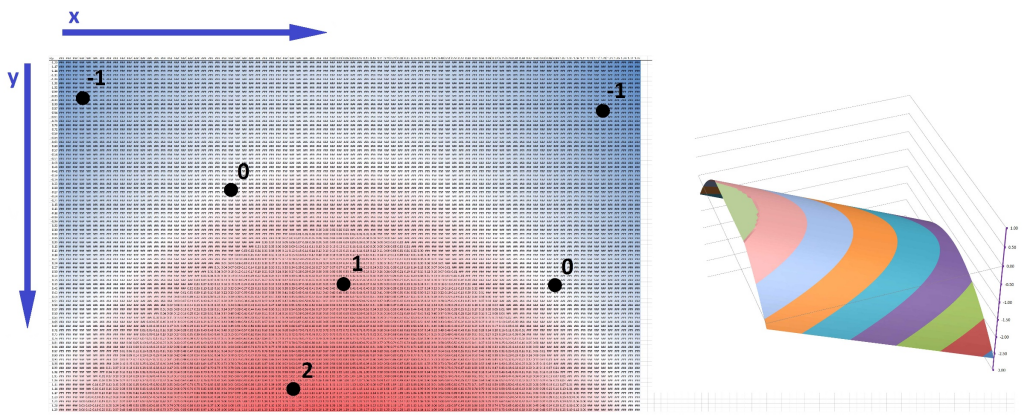
Each hyperbola seems to consist of *two* branches. Justify this observation.

Example 2.8.12: parabolas

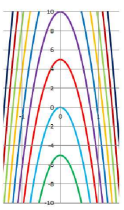
We color the cells of an array according to the values of

$$y - x^2 = k$$

plotted for various k's:



This is what the graphs of these relations look like plotted for various k's; they are curves called *parabolas* (Chapter 4):



They are the graphs of the functions that come from solving the equation for y:

$$y = x^2 + k\,.$$

Exercise 2.8.13

What if we solve for x instead?

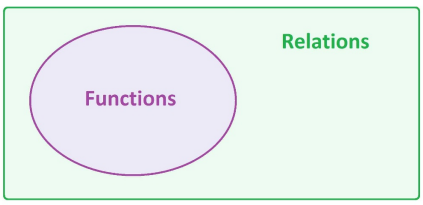
Exercise 2.8.14

Visualize the relation $2x^2 + y^2 = 3$.

Even with the help of a computer, trying to find every point on the whole (x, y)-plane that satisfies a given relation is like looking for a needle in a haystack. In contrast, functions produce the "allowed" pairs (x, y)

automatically, without needing to test each of them. Simply plug in a value, x, and the function will give you its mate, y.

Remember, all functions are relations but not all relations are functions:



This means that what we have said about relations will apply to functions, but we will be able to say much more about the latter.

2.9. A function as a black box

Functions are *explicit relations*. The two variables are still related to each other, but this relation is now unequal: The input comes first and, therefore, the output is *dependent* on the input. That is why we say that the input is the *independent variable* while the output is the *dependent variable*.

A function is a *black box*; something comes in and something comes out as a result, like this:

$$\text{input} \quad \rightarrow \quad \boxed{\blacksquare\blacksquare\blacksquare} \quad \rightarrow \quad \text{output}$$

The only law is the following:

- The same input must produce the same output.

For example, a vending machine will provide you with the item the code of which you have entered (if sufficient funds are inserted).

In the case of numerical functions, both are numbers. The black box metaphor suggests that while some computation happens inside the box, what it is exactly may be unknown:

input		function		output
income	$\rightarrow$	IRS	$\rightarrow$	tax bill

How things happen might be even unimportant; what's important is the rule a function has to follow: one y for each x. For example, if you don't know how this function is computed, you can ask someone to do it for you:

input		function		output
x	$\rightarrow$	cos	$\rightarrow$	y

If we are able to peek inside, we might see something very complex or something very simple:

$$\text{input} \quad \rightarrow \quad \boxed{\text{multiply by 3}} \quad \rightarrow \quad \text{output}$$

Function is what function does! It may be simply a *sequence of instructions*.

Example 2.9.1: flowcharts represent functions

For example, for a given input x, we do the following consecutively:

- add 3,
- multiply by 2, and then
- square.

Such a procedure can be conveniently visualized with a "flowchart":



If the input is $x = 1$, we acquire three more numbers in this order:

$$1 \;\mapsto\; 1+3=4 \;\mapsto\; 4\cdot 2=8 \;\mapsto\; 8^2=64$$

Here is the algebra of what is going on inside of each of the boxes:

$$x \;\to\; \boxed{x+3} \;\to\; y \;\to\; \boxed{y\cdot 2} \;\to\; z \;\to\; \boxed{z^2} \;\to\; u$$

We have introduced *intermediate variables* for reference. Note how the names of the variables match; we, therefore, can proceed to the next step. A sequence of algebraic steps of this process is as follows:

$$\begin{aligned} x \;\to\; x+3 &= y \\ \to\; y\cdot 2 &= z \\ \to\; z^2 &= u \end{aligned}$$

It can also be called an *algorithmic representation*.

Exercise 2.9.2

Describe the function that computes a sales tax of 5%.

Exercise 2.9.3

Describe the function that computes a discount of 10%.

Warning!

Such a sequence of commands might need forks in order to represent more complex functions.

Thus, we represent a function diagrammatically as a box that processes the input and produces the output:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ x & \to & \boxed{f} & \to & y \end{array}$$

Here, f is *the name of the function* (in fact, "f" stands for "function"). In this example, the function is *unspecified*. We make it *specific* by describing how it works. Some specific functions are given specific names made of letters and symbols, such as these:

- $\sqrt{(\quad)}$ for the square root
- $\exp(\quad)$ or $e^{(\quad)}$ for the exponential function
- $\sin(\quad)$ for the sine, etc.

Numerical functions come from many sources and can be expressed in different forms:

- a list of instructions (an algorithm)
- an algebraic formula
- a list of pairs of inputs and outputs
- a graph
- a transformation

We will be transitioning from one to the next as needed with the exception of the last item, which is postponed until the next chapter.

An algorithm is commonly a list of instructions given to a computer, i.e., a *program*. It may be preferable to have a function to be handled by a person represented in the form of a *formula*. The person may appreciate a more compact form that allows to notice patterns, simplify, and further manipulate the function.

An *algebraic representation* is exemplified by $y = x^2$. In order to properly introduce this as a function, we give it a name, say f, and write:

$$f(x) = x^2\,.$$

Let's examine this notation:

Variables of function

	y	$=$	f	$($	x	$)$	$=$	x^2
	↑		↑		↑			↑
name:	dependent variable		function		independent variable			independent variable

The letters are all just *names*! The choices for these names are mostly arbitrary. They have to vary when there is more than just one function present, for example:

	z	$=$	g	$($	t	$)$	$=$	$t+5$
	↑		↑		↑			↑
name:	dependent variable		function		independent variable			independent variable

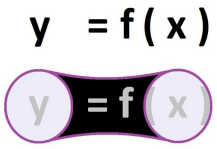
Warning!

It is often acceptable (or even preferable) to omit the name of the function and concentrate on the variables, as we did in the last example.

Thus, the independent variable is the input, and the dependent variable is the output. When the independent variable is made specific, so is the dependent variable, via the *substitution*:

f	$($	3	$)$	$=$	3^2		
↑		↑			↑		
function		input			output		

We can think of this notation as a "black funnel":



Here x enters through the funnel and then (after processing) y appears from the other end. With the same effect, we can use a blank box as an entry gate instead of x:

f	$($	$\square$	$)$	$=$	$\square^2$		
		↑			↑		
		insert input			insert input		

Example 2.9.4: plug in values

For a more complex function, there may be several boxes, but the idea remains the same: Insert the input value in all of these boxes. For example, this function:

$$f\left(x\right)=\frac{2x^2-3x+7}{x^3+2x+1},$$

can be understood and evaluated via this diagram:

$$f\left(\square\right)=\frac{2\square^2-3\square+7}{\square^3+2\square+1}.$$

For instance, to find $f(3)$ we just insert 3 at each of these windows:

$$f\left(\boxed{3}\right)=\frac{2\boxed{3}^2-3\boxed{3}+7}{\boxed{3}^3+2\boxed{3}+1}.$$

It is as if the opening of the funnel is split and the value of x is copied to several tubes that feed this value to these locations within the formula.

Warning!

When substituting, use parentheses generously:

$$x-5=(x-5),\ -1=(-1),\quad \text{etc.}$$

Example 2.9.5: substitution of values

If we carelessly substitute expressions that are more complex, errors are inevitable. For example, consider the same function,

$$f\left(x\right)=\frac{2x^2-3x+7}{x^3+2x+1},$$

as above but evaluate it at $x=-2$. "Plugging in" might produce gibberish:

$$f\left(-2\right)=\frac{2-2^2-3-2+7}{-2^3+2-2+1}.$$

One may consider a slightly different diagram:

$$f\boxed{(x)}=\frac{2\boxed{(x)}^2-3\boxed{(x)}+7}{\boxed{(x)}^3+2\boxed{(x)}+1}.$$

In other words:

- ▶ Don't substitute x, substitute (x).

Just as the function notation suggests! Then, we have

$$f\left(-2\right)=\frac{2(-2)^2-3(-2)+7}{(-2)^3+2(-2)+1}.$$

We might have to face even more complex substitutions in the future:

$$f\left(t+5\right)=\frac{2(t+5)^2-3(t+5)+7}{(t+5)^3+2(t+5)+1}.$$

In summary,

▶ "x" in a formula serves as a *placeholder* for: numbers, variables, and whole functions.

Exercise 2.9.6

Provide a formula for the new function $f(z^2)$ made from the function f above.

Example 2.9.7: function as sequence of steps

Let's take the function from the beginning of the section; it requires several stages:

$$y = x + 3, \quad z = y \cdot 2, \quad u = z^2 .$$

They can be written as follows:

$$\square \;\to\; \boxed{\square + 3} \;\to\; \boxed{\square \cdot 2} \;\to\; \boxed{\square^2} \;\to\; \square$$

For example, we compute its output for the input $x = 2$ in three consecutive steps:

$$2 \;\mapsto\; 2 + 3 = 5 \;\mapsto\; 5 \cdot 2 = 10 \;\mapsto\; 10^2 \;\mapsto\; 100$$

Example 2.9.8: decomposition of function

Consider this formula:

$$f(x) = \sqrt{x^2 - 3} + 5 .$$

To represent this function as a list of instructions, we just read the formula starting with x:

$$\text{input} \;\to\; \boxed{\text{square}} \;\to\; \boxed{\text{subtract 3}} \;\to\; \boxed{\text{take square root}} \;\to\; \boxed{\text{add 5}} \;\to\; \text{output}$$

We read outside out!

Exercise 2.9.9

Represent this function as a list of instructions:

$$f(x) = \left(\sqrt{x} + 2\right)^3 .$$

Example 2.9.10: formula from flowchart

Conversely, a diagram can be converted to a single formula. Let's take a flowchart from one of the examples above and *substitute* the variables along the arrows eliminating the intermediate variables one by one:

$$\begin{array}{ccccccc} x \;\to\; x+3 & = & y & & y \text{ is gone.} & & z \text{ is gone.} \\ & & \downarrow & & x \;\to & & x \;\to \\ & & y \cdot 2 & = & z & (x+3)\cdot 2 \;=\; z & \\ & & & & \downarrow & \downarrow & \\ & & & & z^2 \;=\; u & z^2 \;=\; u & \to \left((x+3)\cdot 2\right)^2 \;=\; u \end{array}$$

Therefore, our function is given by the following single formula:

$$u = f(x) = \left((x + 3) \cdot 2\right)^2 .$$

Note that to recover the operations from the formula, we just read it *from inside out*:

$$x$$
$$x+3$$
$$(x+3)$$
$$(x+3)\cdot 2$$
$$\big((x+3)\cdot 2\big)$$
$$\big((x+3)\cdot 2\big)^2$$

Exercise 2.9.11

Find a formula for the following function:

input $\rightarrow$ divide by 2 $\rightarrow$ take its reciprocal $\rightarrow$ subtract 1 $\rightarrow$ output

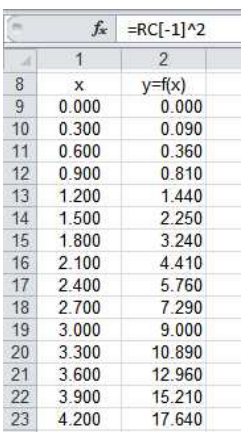
A function can also be represented by a *list of pairs of inputs and outputs.*

This list is a table with two columns, for x and y:

x	$y=f(x)$
0	1
1	3
2	4
3	0
4	2
...	...

This may be called a *numerical representation* of a function as the list contains only numbers. Any list like this would do as long as there are no repetitions in the x-column!

To create larger lists, one uses a spreadsheet. Each value in the y-column is computed from the corresponding value in the x-column via some algebraic formula:



For example, for $y=x^2$, we have in the y-column the following spreadsheet formula:

```
=RC[-1]^2
```

It refers to the value located in: same row, previous column.

Furthermore, it is even possible that a function is pure *data* and there is no formula! One can imagine, for example, that the table has come from a measuring device (say, a thermometer) that takes readings at equal intervals of time.

Even though the data in the list represents the same function as the formula above, we can see that there are gaps in the data. We can't tell, for example, what 1.5^2 is or what 100^2 is. Thus, our algebraic representation is complete, but the numerical representation given by the list is not. However, this list is *a* function but with a smaller domain than the original.

The advantage of numerical representations is that they have been pre-computed for you so that you can see *patterns*; for example, with x increasing we see that:

- y is also increasing, and furthermore,
- y grows faster and faster.

If the last observation is hard to see in the data, we either produce more data – such as compute the difference of the sequence – or visualize the data that we do have.

We can use the list data to plot points, which leads us to the *graphical representation* of functions. Below, the definition we have used for relations is repeated but this one will be even more widely used.

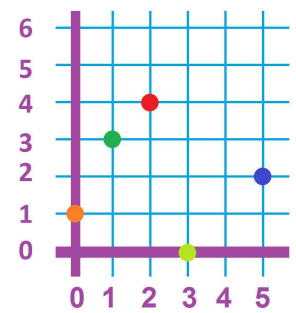
Definition 2.9.12: graph of function

The *graph* of a numerical function $y = f(x)$ is the set of points in the xy-plane that satisfy $y = f(x)$. In other words, it is the following set:

$$\{(x, y) : \ y = f(x)\}.$$

For example, we can plot the above data (just the points that have been provided):

x	$y = f(x)$
0	1
1	3
2	4
3	0
4	2



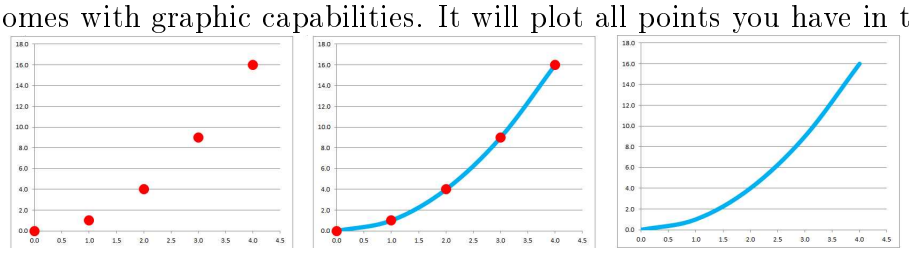
Warning!

We will speak of "a graph", or "graphs", when we deal with the graph of some *function*.

Example 2.9.13: plotting points

A spreadsheet software comes with graphic capabilities. It will plot all points you have in the list:

x	y=f(x)
0.000	0.000
1.000	1.000
2.000	4.000
3.000	9.000
4.000	16.000

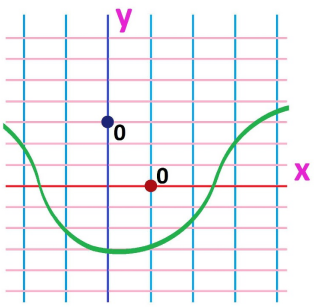


It can also automatically add a curve connecting these points.

Warning!

The first plot is the truth; the rest is a guess.

Note that when x and y represent two variables that have nothing to do with each other – such as time and location – neither do the two axes. In that case, neither the unit lengths nor the locations of the origins have to match:



A *transformation* takes the domain X, a subset of the real line, transforms it according to the function (shift, stretch, flip, etc.), and places the result on the codomain Y. It is discussed in the next chapter.

An *algorithm* is a verbal representation of a function. It may contain no explicit algebra. Instead, it tells us how to get a certain output given any input. For example,

- Question: How do we get from x to y?
- Answer: Let y be equal to the square of x.

This representation, too, conveys a *compete* information about the function.

Example 2.9.14: flowchart from formula

Describe what this function does:

$$f(x) = \frac{x^2+1}{x^2-1},$$

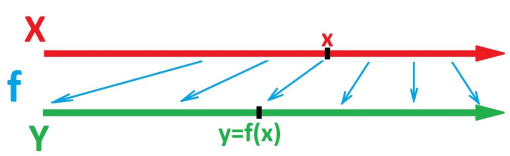
verbally:

- Step 1: Multiply x by itself, call it y.
- Step 2: Add 1 to y, call it z.
- Step 3: Subtract 1 from y, call it u.
- Step 4: Divide z by u.

There is a *fork* in the diagram:

$$x \;\to\; x^2 = y \to \begin{array}{ll} \nearrow & y+1=z \\ \searrow & y-1=u \end{array} \begin{array}{l} \searrow \\ \nearrow \end{array} z/u = w$$

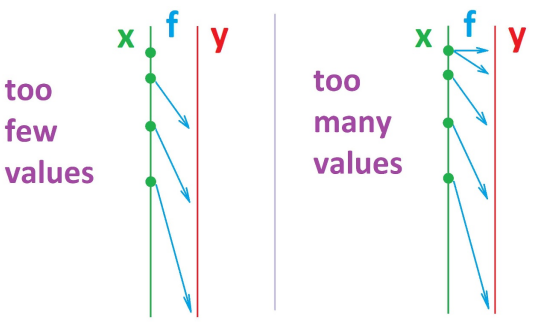
A (numerical) *function* is a rule or procedure f that assigns to any number x in a set X, called the *set of inputs* or the *domain*, one number y in another set of real numbers Y, called the *set of outputs* or the *codomain* of f.



In other words,

1. each x in X has a counterpart in Y, and
2. there is only one such counterpart.

This rule can be violated when there are too few or too many arrows for a given x:



Then this is *not a function*. It is OK, however, to have too few or too many arrows for a given y!

Next, let's revisit the rule – how to get y from x – that defines a function. It must satisfy:

- *There is only one y for each x.*

Let's illustrate how the rule might visibly fail for each of these four ways to represent a function.

Example 2.9.15: algebraic representation

In the following very common way to present a formula, there are two outputs for the same input (unless $x = 0$):

$$y = \pm x\,.$$

Not a function!

Example 2.9.16: numerical representation

In the following list of values, the inputs aren't ordered. It is, therefore, possible that the list might contain two rows with the same x-value and different y-values:

x	y
...	...
0	22
...	...
...	...
...	...
0	55
...	...

same! different!

Not a function!

Example 2.9.17: algorithmic representation

In this list of commands, one is either ambiguous or it produces multiple outputs:

- Step 1: ...
- ...
- Step 50: Add today's date to the output of step 49.
- ...
- Step 100: ...

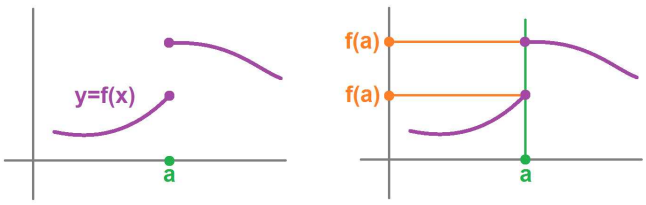
Not a function!

Exercise 2.9.18

Suggest your own examples of how formulas, lists, and algorithms can fail to give us a function.

Example 2.9.19: graphical representation

The following graph has two points – outputs – above $x = a$:



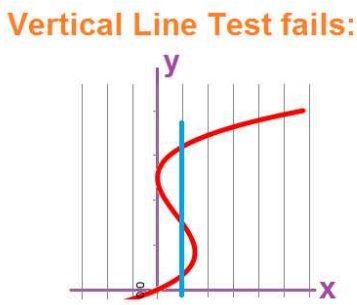
Not a function!

For the graphical representation, all it takes is a glance.

Theorem 2.9.20: Vertical Line Test For Relations

A relation is a function of x if and only if every vertical line crosses the graph at one point or none.

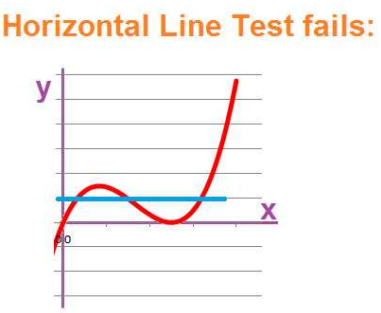
So, every vertical line is a test:



Corollary 2.9.21: Horizontal Line Test For Relations

A relation is a function of y if and only if every horizontal line crosses the graph at one point or none.

So, every horizontal line is a test:



Exercise 2.9.22

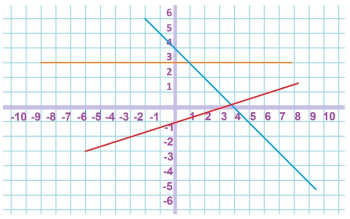
Split either theorem into a statement and its converse.

Example 2.9.23: functions or relations?

Let's apply these tests. A direct examination of the graphs of the relations from the last section produces the following results. First,

- ▶ The relation $Ax + By = C$ is a function of x unless $B = 0$ and is a function of y unless $A = 0$.

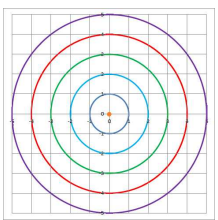
We have a variety of *straight lines*:



Second,

▶ The relation $x^2 + y^2 = k$ is not a function (of x or y).

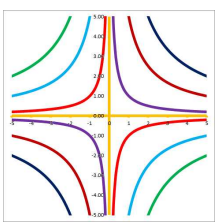
We have a concentric circle for each $k > 0$:



Third,

▶ The relation $xy = k$ is a function of x or y.

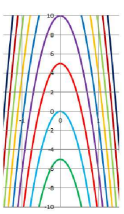
We have a *hyperbola* for each $k \neq 0$:



Fourth,

▶ The relation $x^2 - y = k$ is a function of x but not of y.

We have a shifted *parabola* for each k:



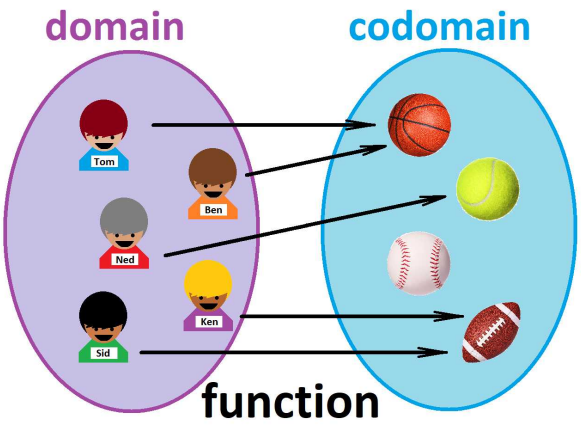
Exercise 2.9.24

Describe the graphs of the relations in the cases missing in the above example:

1. $Ax + By = C$ with $B = 0$ or $A = 0$;
2. $x^2 + y^2 = k$ with $k \leq 0$;
3. $xy = k$ with $k = 0$.

2.10. Give the function a domain...

Recall that a function $F : X \to Y$ is *defined* as a correspondence between two sets, its domain X and its codomain Y:



A function, therefore, is not a function unless these *three* items have been specified:

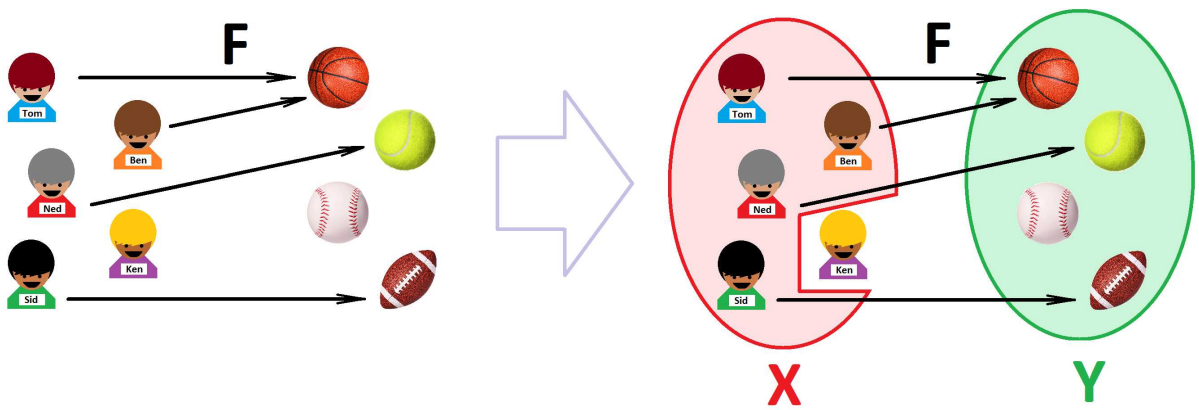
- a domain X,
- a codomain Y, and
- a correspondence F.

Now, what if F is specified but X and Y are not?

In our running example, we may have a collection of arrows, from a boy to a ball. We need to specify X and Y and in such a way that we have a function; i.e.,

- there is exactly one y for each x.

The choices may be obvious when the function is visualized:



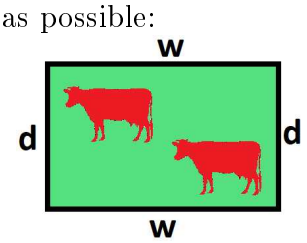
As you can see, Ken has no preferred ball (no arrows) and, therefore, cannot be a part of the domain X.

Warning!

Baseball isn't chosen by anyone (no arrows) and can, but doesn't have to, be excluded from the codomain Y.

Example 2.10.1: maximizing enclosure, continued

Choosing a domain is often a matter of *convenience* and common sense. For example, recall the problem from the beginning of the chapter about cattle enclosure to be made from 100 yards of fencing material with as large an area as possible:



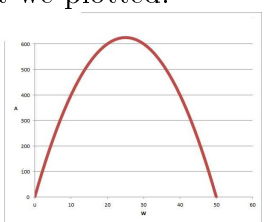
As this was a word problem, it was entirely up to us to decide what quantities, variables, sets, and functions to choose for our solution. In particular, the area of the enclosure was expressed in terms of its width by the formula:

$$a = w(50 - w)\,.$$

Any number w can be processed through this formula. However, we made the implicit choice of the domain:

$$X = [0, 50]\,.$$

Why? The negative widths simply don't make sense! Meanwhile, the widths above 50 produce negative depths. We could have also excluded $w = 0$ and $w = 50$ (on the same grounds), but that would have removed the end-points of the graph that we plotted:



So, we chose to concentrate on those that can be *legitimate solutions* to the problem at hand. In the absence of such a test, we seek another approach.

When we concentrate on *numerical functions*, we may have nothing but a *formula*. Then, we need to make some decisions about the other two attributes.

First, the *codomain*.

The simplest choice is the set of all real numbers. Its advantage is that it is the largest possible and it applies to all functions. So, unless specified otherwise, it is our **convention**:

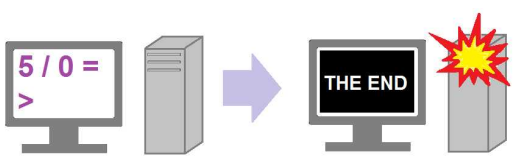
- The codomain of any numerical function is the set of all real numbers $\mathbf{R}$.

In other words, every numerical function is assumed to be one of these:

$$f : X \to \mathbf{R}$$

Now, the *domain*.

Imagine that our function is a list of instructions (an algorithm) and it is used to create a computer program. In this case, x, the input, passes through a black box and out comes y. But if our algorithm requires the computer to divide by x and we give it $x = 0$ as an input, it might do exactly that and there will be *trouble*:



So, if we have a function and the domain isn't specified, it's an oversight.

Let's consider how to handle this issue for the different ways to represent a function.

Example 2.10.2: domain from formula

For a function given by a formula, we just need to choose a set of allowable inputs for the function that we already have. Algebraically, we plug various x's into the formula and see if it works. Let's consider:

$$f(x) = \frac{1}{x}\,.$$

It works for all positive and all negative numbers – but let's try $x = 0$. The function fails because $\frac{1}{0}$

is undefined; no output! Since $\frac{1}{x}$ is defined for every $x \neq 0$, we can choose the domain to be *all* these numbers:

$$X = \{x : \ x \neq 0\}\,.$$

Now, what if we take just one half of this:

$$X = (0, \infty)\ ?$$

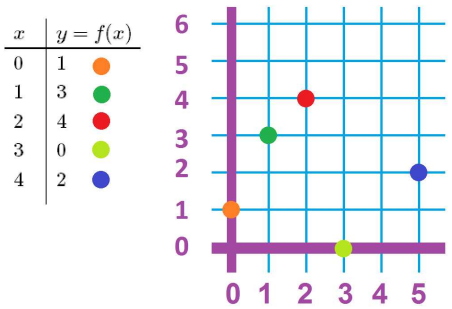
It is also a valid choice for domain. Any set that excludes 0 will do:

$$\{..., -2, -1, 1, 2, ...\},\ [1, 2],\ (-1, 0),\ ...$$

For any such a choice of X, we have a function $f : X \to \mathbf{R}$.

Example 2.10.3: domain from list

If a function is given by a list of inputs and corresponding outputs, the domain can be simply the set of all entries in the first column:

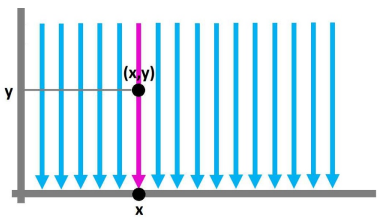


Here we have a function:

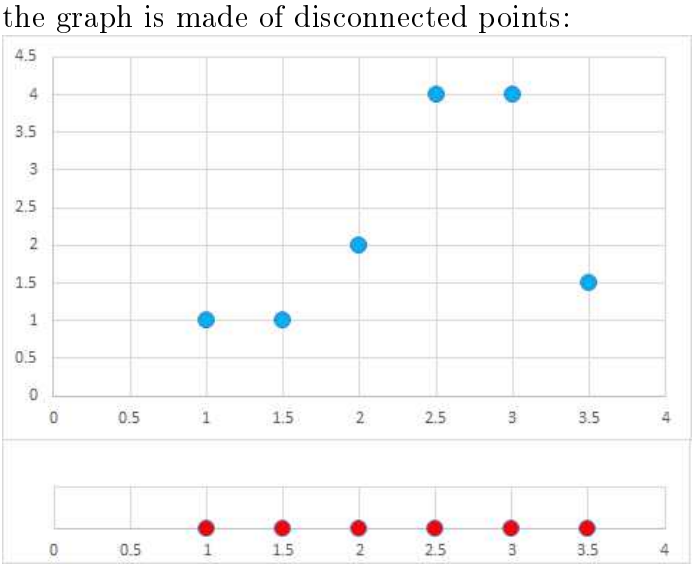
$$f : \{0, 1, 2, 3, 4, 5\} \to \mathbf{R}\,.$$

Example 2.10.4: domain from graph

If the function is given by its graph and nothing else, the domain is found by following each point on the graph to the x-axis in this manner:



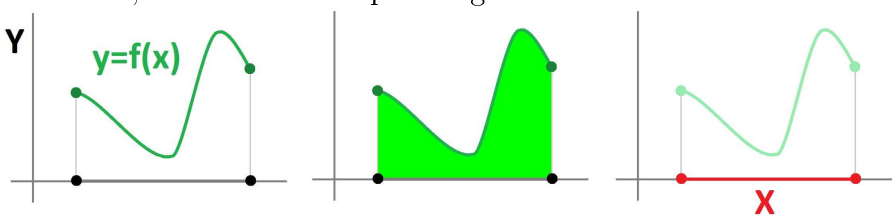
It's especially easy when the graph is made of disconnected points:



We can imagine that balls (blue) were suspended in the air, and then dropped to the ground. These balls (red), as they lie on the x-axis, show the domain points. Here the domain is

$$X = \{1, 1.5, 2, 2.5, 3, 3.5\}\,.$$

When the graph is a curve, we see it as a rope being lowered down on the x-axis:



What is the advantage of one domain over another? The size. A smaller domain is a constraint on what we can do with the function and we, therefore, choose the "largest possible domain". A more precise way to describe this choice is as a certain set.

Definition 2.10.5: implied domain

The *implied domain* (or the natural domain) of a formula or a correspondence is the set of all inputs for which it produces outputs.

In other words, the implied domain of f is the following subset of the real line:

$$X = \{x :\ f(x)\ \text{ makes sense }\}$$

To find this set, we need to look at the formula and answer the question: What can go wrong?

Example 2.10.6: domain from formula: division

find the implied domain of the function:

$$f(x) = \frac{x^2+1}{x^2-1}\,.$$

We need to ensure that the input x doesn't produce a 0 in the denominator. We have the domain presented in the set-building notation:

$$X = \{x :\ x^2 - 1 \neq 0\}\,.$$

The problem is solved, to a degree. The second stage could be to simplify and visualize this set. To find those x's that are *not* in the domain, we need to solve the equation:

$$x^2 - 1 = 0\,.$$

It follows that $x^2 = 1$. Thus $x = -1$ and $x = +1$. The function is defined by all values except ± 1. Then, we have a simplified representation of the implied domain:

$$X = \{x :\ x \neq \pm 1\}\,.$$

So, while technically a function isn't a function without a domain, a formula carries with it its own domain!

Any two numbers can be added, subtracted, or multiplied, but division has an exception. What else can go wrong? Square roots of negative numbers are undefined (because the square of two negative numbers is always positive).

Example 2.10.7: domain from formula: roots

Consider

$$f(x) = \sqrt{x-1}\,.$$

To find its domain, we need to make sure that the input of the square root isn't negative. What we *don't* want to have is written simply as

$$x - 1 < 0\,.$$

Solving this *inequality* gives us:

$$x < 1\,.$$

If in the last example we took the solution set of an equation and excluded it from the domain, we now do that with the solution set of this inequality. Therefore, the implied domain is

$$X = \{x :\ x \geq 1\} = [1, \infty)\,.$$

So, unless specified otherwise, this will be our **convention**:

- *The domain of any numerical function is its implied domain.*

There are only a couple of "problematic" algebraic operations, for now:

1. division (possibly by 0), and
2. even degree roots (of possibly negative numbers).

The problem of finding the domain is solved by the following methods, respectively:

1. Set the denominator equal to 0 and solve the equation.
2. Set the expression inside the root less than zero and solve the inequality.

In either case, we produce the solution set that is then excluded from the set of real numbers. The result is the implied domain. How these two problems are solved is further discussed in Chapter 5.

Exercise 2.10.8

Show that the appearance of a division or a square root in the formula of the function doesn't always cause the domain to lose points:

$$f(x) = \frac{1}{x^2+1}, \quad g(x) = \sqrt{x^2+1}\,.$$

Example 2.10.9: multiple conditions

Find the implied domain of the function given by the formula:

$$f(x) = \frac{\sqrt{x-1}}{x+1}\,.$$

We see *two* problematic operations! We take case of them separately, at first. For the numerator to be defined, we require:

$$x - 1 \geq 0\,.$$

For the denominator to make sense, we require:

$$x + 1 \neq 0\,.$$

Now, the function is defined only when both of the conditions are satisfied. Therefore, the implied domain is the following:

$$X = \{x :\ x - 1 \geq 0 \quad \texttt{AND} \quad x + 1 \neq 0\}\,.$$

Exercise 2.10.10

Find the implied domain of the function given by the formula:

$$f(x) = \sqrt{1+x} + \sqrt{1-x}\,.$$

Example 2.10.11: domain for program

What if the function is given by a list of instructions? For example, let's find the implied domain of the following function:

$$\text{input} \;\to\; \boxed{\text{multiply by 2}} \;\to\; \boxed{\text{add 2}} \;\to\; \boxed{\text{take the reciprocal}} \;\to\; \boxed{\text{subtract 55}} \;\to \ldots \to \text{output}$$

Step 3 is the problem! Through trial and error, we can discover that input $x = -1$ causes an "explosion":

$$-1 \;\to\; (-1)\cdot 2 = -2 \;\to\; (-1)+2 = 0 \;\to\; \frac{1}{0} = \otimes$$

Let's make sure that there are no other problems. The input of step 3 can't be 0. But how do we make sure that this won't happen? First, the input of step 3 is the output of step 2. Therefore, its output can't be 0. How do we ensure that? The input of step 2 can't be -2! Second, the output of step 1 can't be -2 and, therefore, its input can't be -1. This is the only number we don't allow as an input of the whole function. In other words, we *trace back* $z = 0$ (solving equations along the way) to the corresponding value of x:

$$x \;\to\; x\cdot 2 = y \;\to\; y+2 = z \;\to\; \frac{1}{z} = u \;\to\; u - 55 = w \ldots$$

$$x \;\to\; x\cdot 2 = y \;\to\; y+2 = z \;\to\; z = 0 \;\otimes$$

$$x \;\to\; x\cdot 2 = y \;\to\; y+2 = z = 0 \;\otimes$$

$$x \;\to\; x\cdot 2 = y = -2 \;\otimes$$

$$x = -1 \;\otimes$$

So, the implied domain is

$$X = \{x :\; x \neq -1\}\,.$$

Note that the operations that come *after* division in step 3 won't change our decision about excluding $x = -1$ from the domain but might produce more exclusions.

Exercise 2.10.12

What if the next (and the last) operation is the square root? Find the implied domain of the function.

Exercise 2.10.13

Find the implied domain of the following function:

$$\text{input} \;\to\; \boxed{\text{take the square root}} \;\to\; \boxed{\text{add 3}} \;\to\; \boxed{\text{take the reciprocal}} \;\to \text{output}$$

Exercise 2.10.14

Find the implied domain of the following function:

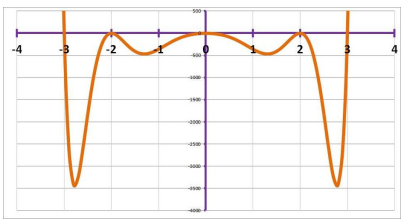
$$\text{input} \;\to\; \boxed{\text{add 3}} \;\to\; \boxed{\text{square it}} \;\to\; \boxed{\text{add 2}} \;\to\; \boxed{\text{divide by 0}} \;\to\; \text{output}$$

In summary, when we face a function without a domain or codomain, it is *implied* that this is a function with the following features:

The domain	is	the implied domain.
The codomain	is	all real numbers.

2.11. The graph of a function

A function may be given to us in the form of a list, a formula, or a list of instructions. Those deal with the function one input at a time. This is why one will find it hard to discern *patterns* that may be hidden in the function:



Graphs provide a way to have a bird's eye view of the function.

Recall that the *graph* of a relation is the set of points in the xy-plane that satisfy the relation. In the case of a function then, the graph of a function $y = f(x)$ is the set of points in the xy-plane that satisfy $y = f(x)$. In other words, it is the set of all possible points on the plane in the form $(x, f(x))$:

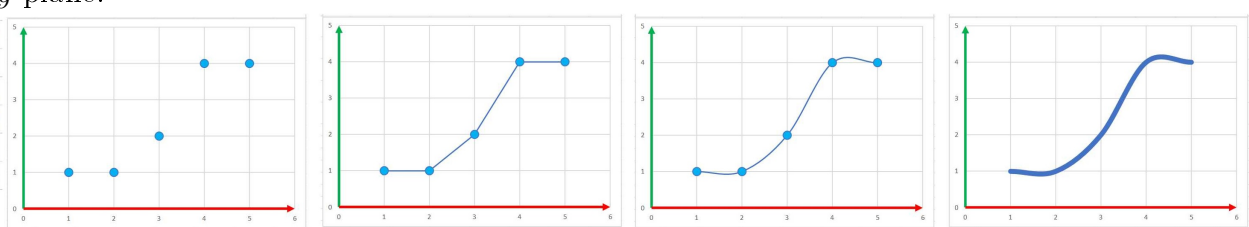
$$\text{graph of } f \;=\; \{(x, y) :\; y = f(x)\}$$

We use the set-building notation again!

Example 2.11.1: function from list

Suppose we have a function represented by a list of pairs of values, x and y. We can use the data in list to plot points, which leads us to its graphical representation. The list and the first graph below have been seen before. We simply treat each of the rows of the list as the two coordinates of a point on the xy-plane:

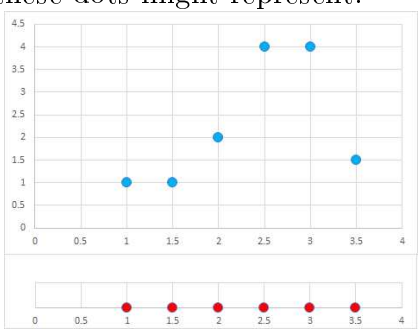
X	Y
1	1
2	1
3	2
4	4
5	4



The domain of the function is just these five values of x. Furthermore, we may try to *extrapolate* the data to the whole interval. The first graph is data; the rest are just guesses.

Exercise 2.11.2

Guess the shape of the curve that these dots might represent:

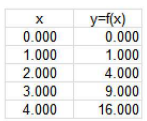


Example 2.11.3: function from formula

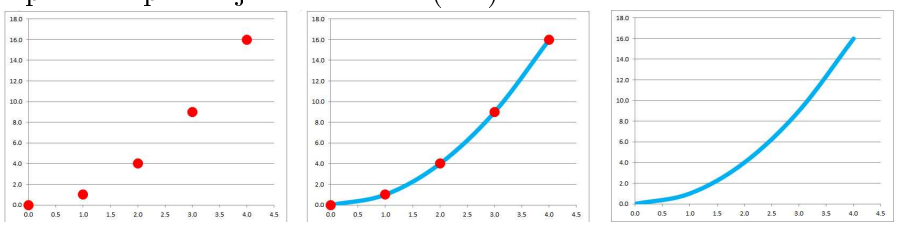
What if the function is given by a formula, say, $y = f(x) = x^2$? We still have to build a list! We make a table with two columns, for x and y, and then fill it using the formula one row at a time:

x	$y = x^2$
0	0
1	1
2	4
3	9
4	16

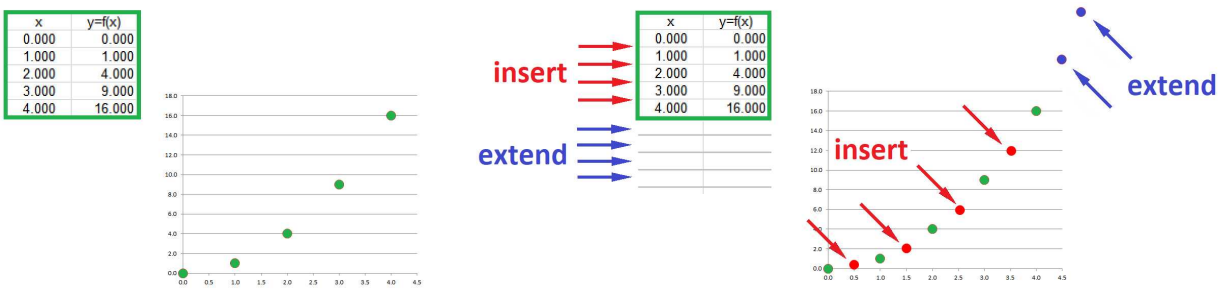
With a list ready, we then plot the points just as before (left):



x	y=f(x)
0.000	0.000
1.000	1.000
2.000	4.000
3.000	9.000
4.000	16.000



Note that there are gaps in the data; we can't tell, for example, what 1.5^2 is. The data also don't go far enough; we can't tell what 100^2 is. Thus, our algebraic representation – the formula – is complete but the numerical representation given by the list is not. The graph is, therefore, also just a *partial representation*! We can *guess* what happens between the points or we can let the spreadsheet do it and automatically add a curve connecting the points (above) or add more values to the list:

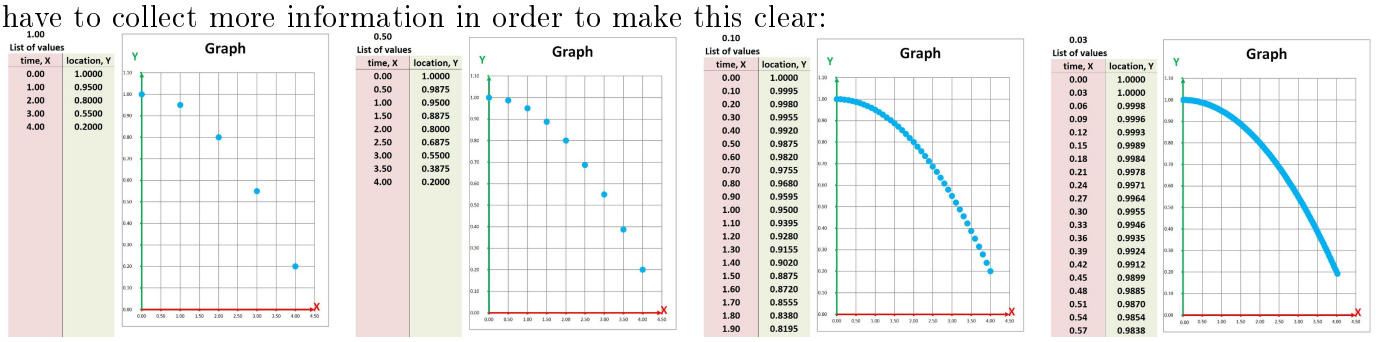


Exercise 2.11.4

Plot the graph of the function given by this list of instructions: (i) add 3, (ii) divide by 2, (iii) square the outcome.

Example 2.11.5: continuum?

If it is known that our function is just a snapshot of a "continuous" process, such as motion, we may have to collect more information in order to make this clear:

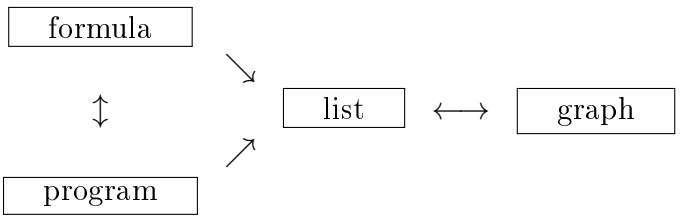


We, for example, may look at the odometer every minute, or every second, etc., instead of every hour. The infinite divisibility of the real line (Chapter 1) allows us to produce sets of points on the plane with denser and denser patterns. We imagine that at the end of this process we will have an actual curve. This is not a kind of curve that is made of marbles placed close together, but a rope. What happens "at the end" is studied by calculus (Chapter 2DC-2).

Warning!

The graph of a function isn't the function; it is only a visualization of some of its data.

In summary, this is how graphs appear:

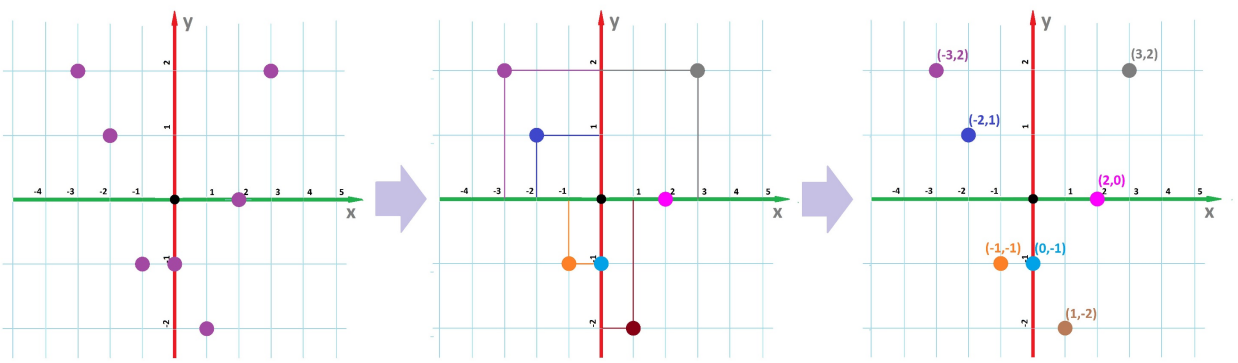


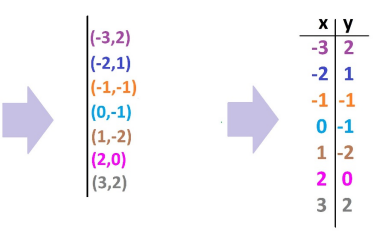
The double arrows are reversible.

How can we reverse the direction of the last arrow? How can we find the list of values of the function if we only have its graph? It's feasible when the graph is small.

Example 2.11.6: from function to list

We go from point to point and find the coordinates of each. Then we put these pairs of points one under the other in a list:

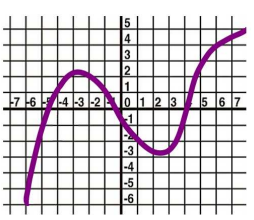




The domain of the new function is automatically constructed in the first column of the list.

Exercise 2.11.7

Create a list of 10 values from this graph:



Of course, if we start with a list and plot the graph, then the list built from the graph, as above, is the original list! However, we can't hope to recover an infinitely long list nor a formula.

Warning!

Even though we should normally refer to it as "the graph of a function", we may informally refer to

a curve that passes the Vertical Line Test as "a graph".

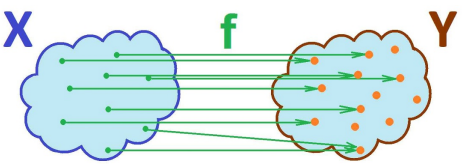
Can we ever treat a *graph as a function*?

Let's recall the idea of function as a *black box* that processes the input and produces the output:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ x & \to & \boxed{f} & \to & y \end{array}$$

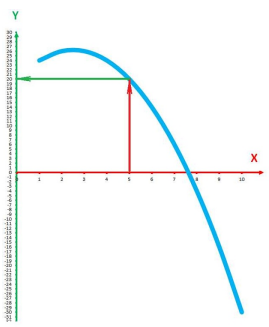
$$\begin{array}{ccccccc} f & (& 3 &) & = & 3^2 \\ \uparrow & & \uparrow & & & \uparrow \\ \text{function} & & \text{input} & & & \text{output} \end{array}$$

Now, suppose we have a graph and nothing else. Where in the graph is that black box? Let's find the arrows from x's to y's that we used to illustrate functions in the beginning of the chapter:

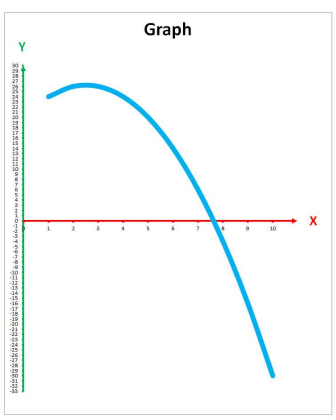


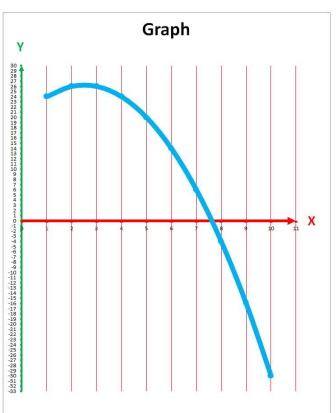
Suppose a *graph* is given to us on a piece of paper. Let's build a black box for the function, f, it represents. For each x, we need to find y using nothing but the graph. How?

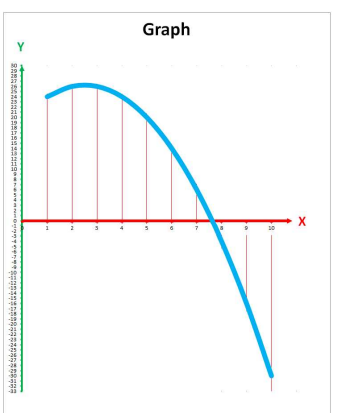
We just reverse the process of building the graph from a list of values. What we do for a single input is to draw a red vertical line until it crosses the graph, and then from that point, we draw a green horizontal line until it crosses the y-axis, as follows:

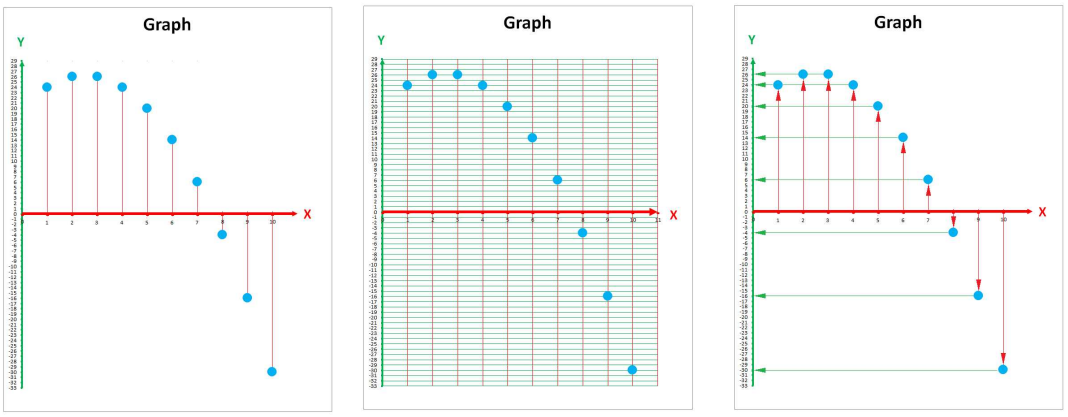


We do this for as many locations on the x-axis as possible. This is the totality of inputs and outputs connected by arrows:









These arrows give us a visualization of our newly-built function. We can, furthermore, represent this function as a *list of instructions* with the output of each step becoming the input of the next step:

1. A number x is an *input*, a number.
2. Plot the point with *that* coordinate on the x-axis of the xy-plane.
3. Draw through *that* point a vertical line in the xy-plane.
4. Find the point of intersection of *that* line with the graph.
5. Draw through *that* point a horizontal line in the xy-plane.
6. Find the point of intersection of *that* line with the y-axis.
7. Find the coordinate of *that* point.
8. This number, $y = f(x)$, is the *output*.

Just as in the last section, we can examine the list to find what can go *wrong* with the procedure. There are two possibilities:

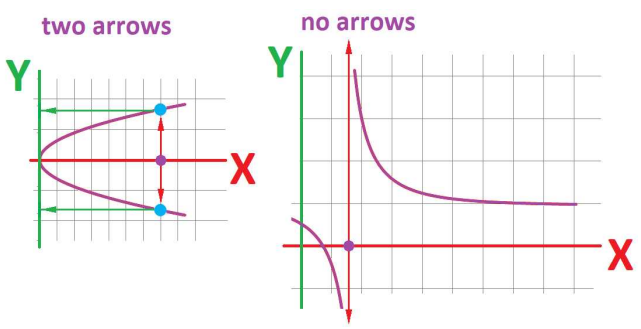
- Too few y's for a given x (none).
- Too many y's for a given x (two or more).

The only step that may cause trouble is #4. First, what if there is no intersection? If there is no point in common between the vertical line (the input of this step) and the graph, then there is no such point (the output of this step). As a result, the function breaks down (just as $f(x) = 1/x$ breaks down if the input is $x = 0$). This means that *that* x is not in the domain of our function! We can see this happening with the graph above for the values of x that are too small or too large. So, the implied domain of this function is

$$\text{domain } = \{x : \quad \text{there is a point on the graph with its } x\text{-coordinate equal to } x\}.$$

Second, what if there are more than one such point of intersection? Then this is simply not a function! We can make it one by removing some of the outputs though.

These are the two ways our algorithm might break down:



Exercise 2.11.8

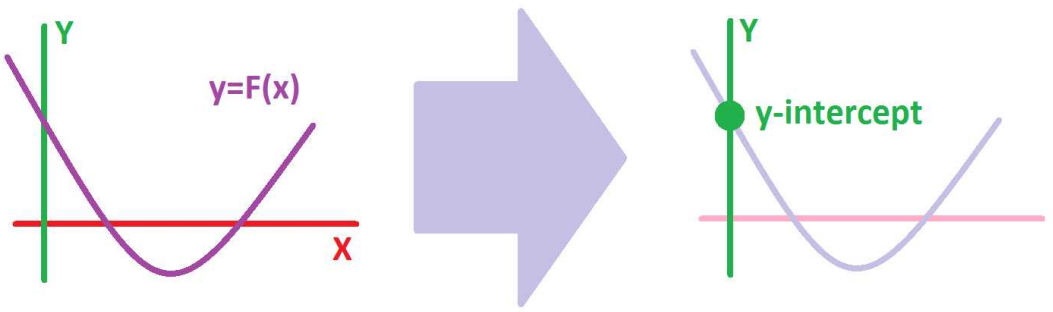
Make the graph on the left into that of a function in three different ways.

Exercise 2.11.9

Write a list of instructions on how to obtain a list of values from a graph.

Graphs, usually, have these two important features.

First is the point of the graph of f that belongs to the y-axis:



Definition 2.11.10: y-intercept

Suppose $y = f(x)$ is a numerical function. Then *the y-intercept* of f is the number y that satisfies $f(0) = y$.

There can be only one (or none, depending on the domain) of these. Finding it amounts to a simple *substitution*.

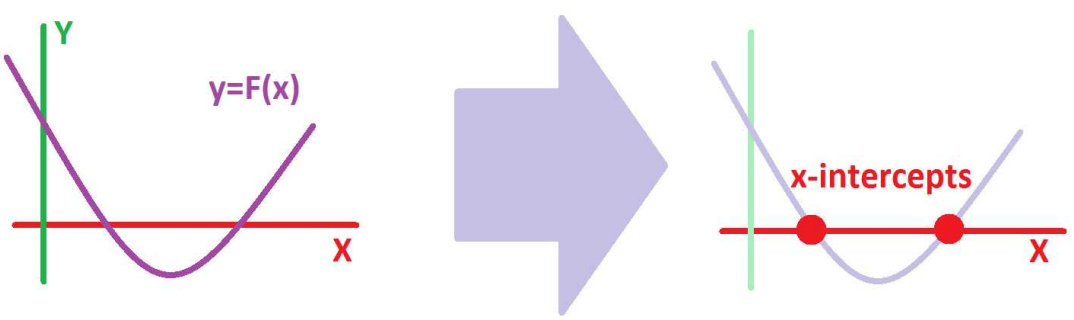
Example 2.11.11: algebra of y-intercept

The y-intercept of $f(x) = x^2 - 1$ is found as follows:

$$y = f(0) = 0^2 - 1 = -1\,.$$

The point $(-2, 0)$ lies on the graph of f.

The matching definition is about the point of the graph of f that belong to the x-axis:



.

Definition 2.11.12: x-intercept

Suppose $y = f(x)$ is a numerical function. Then an *x-intercept* of f is any number x that satisfies: $f(x) = 0$.

There can be one or many depending on the domain of these. Finding them amounts to *solving an equation*.

Example 2.11.13: algebra of x-intercept

The x-intercept of $f(x) = x^2 - 1$ is found by solving the above equation, as follows:

$$f(x) = 0 \implies x^2 - 1 = 0 \implies x = \pm 1\,.$$

The points $(-1, 0)$ and $(1, 0)$ lie on the graph of f.

In other words, these are the x-coordinates of the points where the graph crosses the x-axis.

Exercise 2.11.14

Find the x- and y-intercepts for the graphs in this section.

2.12. Linear functions

The dependence of x on y in a numerical function can be very simple.

However, the simplest kind of function is the one whose output does not change with the input! This is a *constant function*, i.e., it is given by a formula:

$$f(x) = k \quad \text{for each } x,$$

for some predetermined number k. Its implied domain is, of course, $X = (-\infty, \infty)$. Its computation is non-existent; for example, when $k = 3$, we have the following:

$$\text{input} \quad \to \quad \boxed{\text{produce 3}} \quad \to \quad \text{output}$$

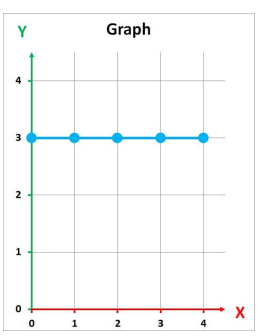
As you can see, the input is thrown away. This is the list of values of this function:

x	$y = f(x)$
0	3
1	3
2	3
3	3
4	3
...	...

Warning!

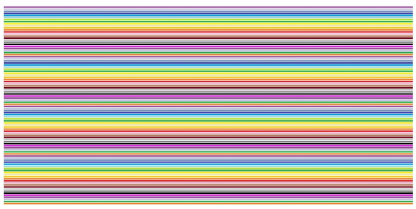
Depending on context, 3 might mean a function.

Plotting a few of these points reveals that the graph is a horizontal line:



Indeed, the corresponding relation is $y = 3$.

This is what the graphs of all constant functions combined look like:



The next simplest function is the one that *does nothing* to the input; i.e., it is given by a formula:

$$f(x) = x\,.$$

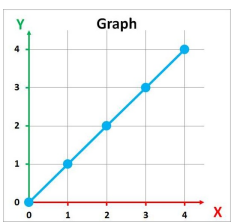
Its implied domain is, of course, $X = (-\infty, \infty)$. Its computation is trivial:

$$\text{input} \quad \to \quad \boxed{\text{pass it}} \quad \to \quad \text{output}$$

This time, the input isn't thrown away but there was still no algebra needed. This is its list of values:

x	$y = f(x)$
0	0
1	1
2	2
3	3
4	4
...	...

Plotting a few of these points reveals that the graph is a 45-degree line:



Indeed, the relation is $y = x$.

Warning!

If we say that y *is* x, then the xy-plane should have the same units for the two axes.

Exercise 2.12.1

Plot the graph of a function that represents the location as it depends on time if the speed is one foot per second.

So far, the function require no algebraic operations! Linear functions are at the next level of complexity. They may be "sloped".

Definition 2.12.2: slope

Suppose we have two points in a specified order, A then B, on the xy-plane, then *the slope of the line* from A to B is defined to be

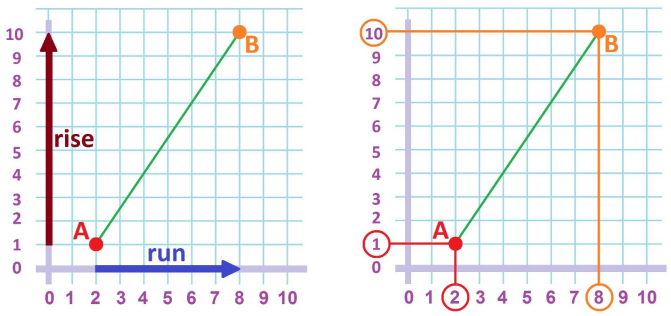
$$\text{slope} \quad = m = \frac{\text{rise}}{\text{run}} = \frac{\text{change of } y}{\text{change of } x}$$

Exercise 2.12.3

Can the rise be zero? Can the run be?

Example 2.12.4: slope

The geometric meaning of the numerator and denominator is seen below:



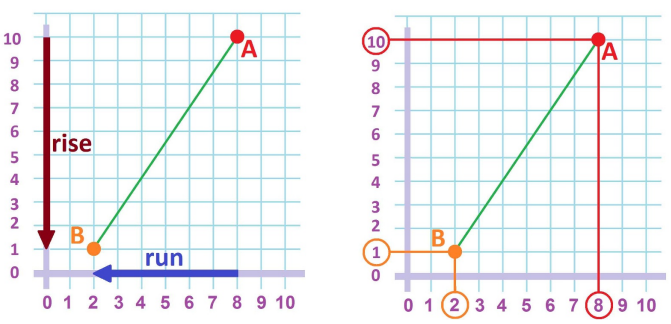
We can just count the number of steps vertically and horizontally (left):

- run $= 6$, and
- rise $= 9$, therefore,
- slope $= \dfrac{9}{6} = \dfrac{3}{2} = 1.5\,.$

Or we can utilize the coordinates of the two points and subtract those of A from those of B (right):

- run $= 8 - 2 = 6$, and
- rise $= 10 - 1 = 9$, therefore,
- slope $= \dfrac{9}{6} = \dfrac{3}{2} = 1.5\,.$

"Rise" and "run" in this context aren't meant to be substitutes for "lengths of these segments" or "distances between those points". In contrast to plain geometry, one or both of them can be negative! In particular, the slope remains unchanged if we reverse the order of the two points: B first, A second:



Indeed:

- run $= 2 - 8 = -6$, and
- rise $= 1 - 10 = -9$, therefore,
- slope $= \dfrac{-9}{-6} = \dfrac{3}{2} = 1.5\,.$

Same slope! It follows that we are studying the *slope of the line* not just that of the two points.

Exercise 2.12.5

Find more pairs of points on the line with slope 1.5.

W utilize the coordinate system to find the slope. Suppose we have two distinct points on a straight line in a specified order, say,

$$A = (x_0, y_0) \quad \text{and} \quad B = (x_1, y_1)\,,$$

then the *slope* of the line they determine is given by the formula:

$$m = \frac{\text{change from } y_0 \text{ to } y_1}{\text{change from } x_0 \text{ to } x_1} = \frac{y_1 - y_0}{x_1 - x_0}$$

It is crucial to know the following:

Theorem 2.12.6: Slope Backwards

The slope from A to B is equal to the slope from B to A.

Proof.

If we reverse the order of the two points – B then A – both numerator and denominator simply flip their signs:

$$\text{change from } y_1 \text{ to } y_0 = -(\text{change from } y_0 \text{ to } y_1)\,,$$

and

$$\text{change from } x_1 \text{ to } x_0 = -(\text{change from } x_0 \text{ to } x_1)\,.$$

But if the numerator and denominator of a fraction flip their signs, the fraction remains intact: $(-a)/(-b) = a/b$. We have for the slope:

$$m = \frac{y_0 - y_1}{x_0 - x_1} = \frac{-(y_1 - y_0)}{-(x_1 - x_0)} = \frac{y_1 - y_0}{x_1 - x_0} = m\,.$$

Warning!

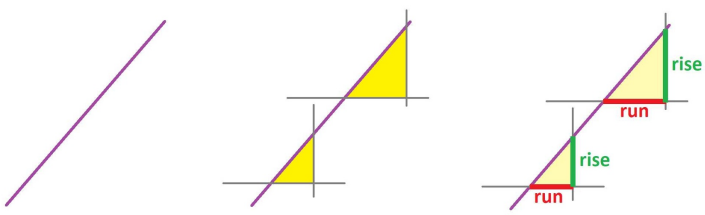
Whether it's A then B, or B then A, it must be the same for both numerator and denominator.

This is what makes a straight line a straight line.

Theorem 2.12.7: Slope From Two Points

Any two points chosen on a straight line produce the same slope.

The theorem says that these two triangles produce the same slope:

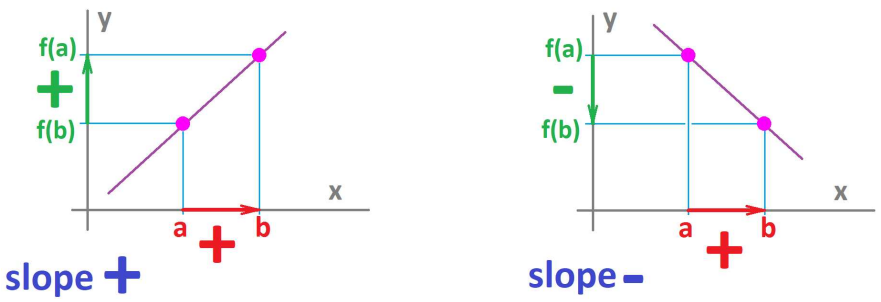


Exercise 2.12.8

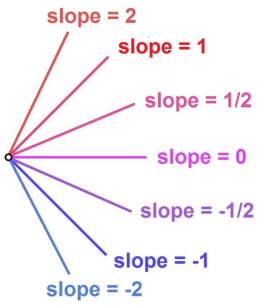
Prove the theorem. Hint: Similar triangles.

What does the slope tell us about the line?

While a positive slope appears when the rise and the run have the same signs, a negative slope appears when the signs are opposite:



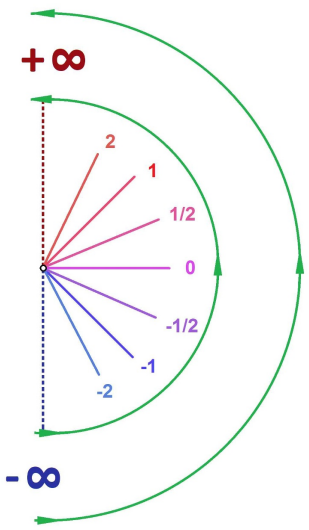
Below we arrange all lines according to their slopes (as if they all start at the origin):



Warning!

Comparing the line with slope $m = 1$ and the line with $m = -2$ suggests that the word "steepness" as a substitute for slope should be used with caution.

It's as if increasing the slope *rotates* the line counterclockwise:

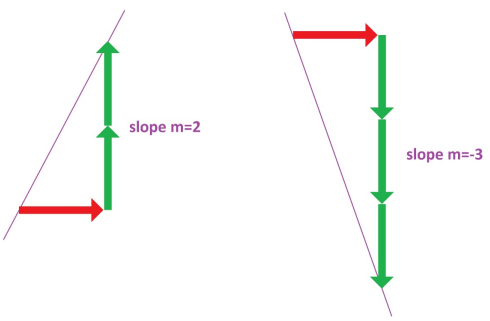


We can see that the scope of possible values of slopes is $(-\infty, +\infty)$.

Warning!

It is impossible to assign a slope to a vertical line. The horizontal like does, however, have a slope; it is zero.

It is clear now that slopes are just another way to look at the *angles* between lines – without trigonometry (Chapter 5). The approach via the slopes, however, is simpler! The slope of m means that as we follow the line, we make a step of 1 unit to the right and then m units up:



In other words, we have in this case:

$$\text{slope} = \frac{\text{rise}}{1}.$$

Of course, this is a down step when $m < 0$.

Warning!

The idea of slope is meaningless without axes as a frame of reference.

Exercise 2.12.9

What happens to the slope of a line drawn on a piece of paper for different choices of the axes?

We exclude the possibility of a vertical line and an infinite slope! This is why we can concentrate on linear *functions* only.

Definition 2.12.10: linear function and polynomial

A *linear function* is a numerical function given by this formula:

$$f(x) = m \cdot x + b$$

for some predetermined numbers m and b. When $m \neq 0$, such a function is called a *linear polynomial*.

Exercise 2.12.11

What is the reason for excluding $m = 0$?

So, the simplest algebra has appeared: addition/subtraction and multiplication by a constant number. They are seen in the function's flow-chart:

$$f: \quad x \quad \to \boxed{\text{multiply by } m} \to \boxed{\text{add } b} \to \quad y$$

The formula is commonly called the *slope-intercept form* of the linear function:

Slope-intercept form

$$f(x) = \underset{\substack{\uparrow \\ \text{slope}}}{m} \cdot x + \underset{\substack{\uparrow \\ y\text{-intercept}}}{b}$$

The latter is indeed the *y-intercept* of the function as defined in the last section:

$$f(0) = m \cdot 0 + b = b\,.$$

The concept of slope is central in calculus. For example, in similarity with sequences, we notice the following:

- If $m > 0$, then the outputs $y = f(x)$ are increasing as the inputs x are increasing.
- If $m < 0$, then the outputs $y = f(x)$ are decreasing as the inputs x are increasing.
- If $m = 0$, then the outputs $y = f(x)$ remain the same as the inputs x are increasing; i.e., f is a constant function.

Warning!

Even though straight lines remain straight lines if we resize the plot, the "slopes" will appear to change.

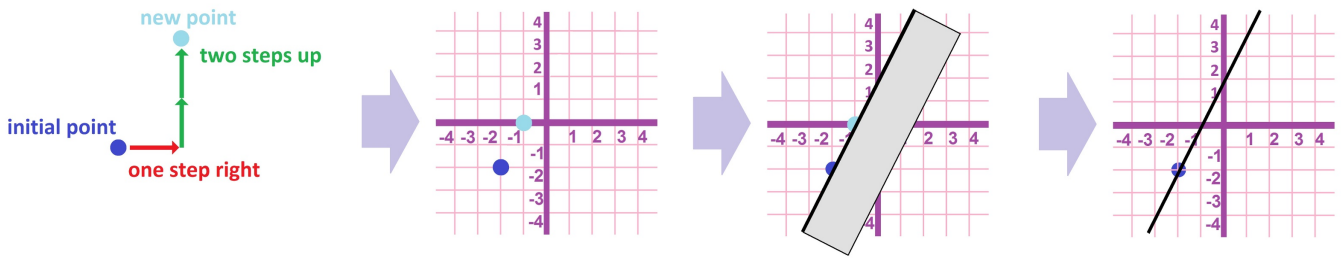
Exercise 2.12.12

Arrange all linear polynomials with the same slope according to their y-intercepts.

The slope gives us the *direction* of the line. That's how the slope-intercept formula, $y = mx + b$, works: We start at the y-intercept, $(0, b)$, and then proceed in the direction provided by the slope, m. In the same manner, we can start at *any* point. Suppose a point is given, say, $A = (x_0, y_0)$. From there, we go as described above: 1 unit right (the run) and m units up (the rise).

Example 2.12.13: plotting a line with a ruler

Let's plot the straight line with slope $m = 2$ through the point $A = (-2, -2)$. From A, we make one step right and two steps up. We have a new point, say B, with coordinates $B = (-1, 0)$ (left):



With a ruler, we draw a line through A and B (right).

Exercise 2.12.14

Plot the straight line with slope $m = -2$ through the point $A = (-1, -1)$. Make up your own parameters and plot the line. Repeat.

Exercise 2.12.15

What is the equation of the line through the points $A = (-1, 2)$ and $B = (2, 1)$?

Example 2.12.16: plotting a line without a ruler

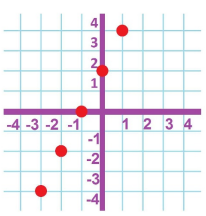
Suppose, again, a point is given, $A = (x_0, y_0)$, and the slope is known to be m. From A, we go 1 unit right and m units up, repeated as many times as necessary:

$$(x_0, y_0) \longrightarrow (x_0 + 1, y_0 + m) \longrightarrow (x_0 + 2, y_0 + 2m) \longrightarrow (x_0 + 3, y_0 + 3m) \longrightarrow \ldots$$

We can move left too:

$$(x_0, y_0) \longrightarrow (x_0 - 1, y_0 - m) \longrightarrow (x_0 - 2, y_0 - 2m) \longrightarrow (x_0 - 3, y_0 - 3m) \longrightarrow \ldots$$

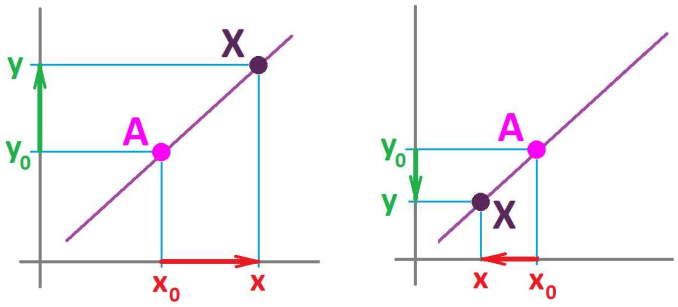
We have a sequence of points forming a line:



Exercise 2.12.17

Plot as many points as possible for the line from $(1, 3)$ and slope -1.

Now, the *algebra*. Suppose we have a specified point $A = (x_0, y_0)$ on our line. Let's consider an arbitrary point $X = (x, y)$ on the line:



The run is $x - x_0$ and the rise is $y - y_0$ (left or right). Therefore, the slope is

$$m = \frac{y - y_0}{x - x_0}.$$

Here x cannot be equal to x_0. To avoid this limitation, we rewrite this formula in such a way that we produce a new and very important way to represent a line.

Theorem 2.12.18: Point-Slope Form of Line

A line with slope m passing through point (x_0, y_0) is given by the following linear relation:

$$y - y_0 = m \cdot (x - x_0)$$

Exercise 2.12.19

What is the difference between the two relations:

$$y - y_0 = m \cdot (x - x_0) \quad \text{and} \quad m = \frac{y - y_0}{x - x_0} \ ?$$

Even though we can solve for y any time we want (and make it a function!), this form is often preferable because of the information it reveals. First, the rise and the run are clearly visible:

Point-slope form of line

$$\begin{array}{ccccc} \text{rise} & = & \text{slope} & \cdot & \text{run} \\ (y - y_0) & = & m & \cdot & (x - x_0) \end{array}$$

Second, the coordinates of the fixed point $A = (x_0, y_0)$ and a variable point $X = (x, y)$ on the graph are visible too:

Point-slope form of line

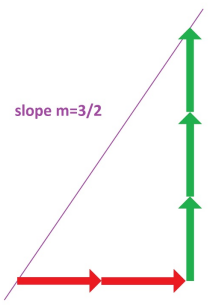
$$\begin{array}{ccccccccc} \text{point } X & & & & & & \text{point } X & & \\ \downarrow & & & & & & \downarrow & & \\ y & - & y_0 & = & m & \cdot & (x & - & x_0) \\ & & \uparrow & & & & & & \uparrow \\ & & \text{point } A & & & & & & \text{point } A \end{array}$$

Exercise 2.12.20

Find the y-intercept from the point-slope form.

Example 2.12.21: plotting incremental motion

What is the slope of the line that follows this path? We make p steps right and q step up as we follow the line:



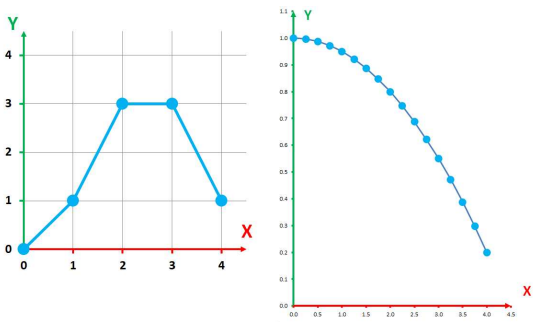
Then the equation becomes:

$$p(y - y_0) = q(x - x_0)\,.$$

The slope is $m = q/p$.

Example 2.12.22: velocity

Setting graphs aside, the slope is the characteristic of a linear function that tells us *how fast* the output is changing relative to the change of the input. An important illustration of this idea was seen early in this chapter when these were location and time, respectively; then the slope is the *velocity*. The examples also showed how the velocity may change incrementally and cause the location to change linearly, interval by interval:



In fact, we might be able to zoom in on a curve and see the same pattern. In light of Chapter 1, we see here two sequences:

$$x_0,\ x_1,\ \ldots \quad \text{and} \quad y_0,\ y_1,\ \ldots$$

The slope of each segment is then the ratio of the corresponding *differences*:

$$m = \frac{\text{difference of } y_n}{\text{difference of } x_n} = \frac{y_{n+1} - y_n}{x_{n+1} - x_n} = \frac{\Delta y}{\Delta x}\,.$$

This expression is called the *difference quotient*. It gives us the velocity when y_n is location and x_n is time. We thus face another, non-geometric, interpretation of the slope, one that will be important throughout calculus: the *rate of change* of the function.

Exercise 2.12.23

Suppose both the domain and the codomain of a linear function are the integers (as in the picture above). What can you say about the slope in this case?

2.13. Algebra creates functions

More complex algebra produces functions with more complex patterns that we will need to discover.

We introduce new algebra into functions building: multiplication of the input by itself.

First, the *square function*:

$$\boxed{f(x) = x^2}$$

What is the difference between multiplication in $3 \cdot x$ and in $x \cdot x$? To begin with, the former is about tripling a quantity of any nature, while the latter may be about the *area* of a square $x \times x$. Second, computing the latter requires – in contrast to most of the recent examples of functions – making a *copy* of the input first:

$$\begin{array}{ccccccc} & & x & \to & \boxed{\text{pass it}} & & \\ x & \to & & & \downarrow & & \\ & & x & \to & \boxed{\text{multiply by } x} & \to & y \end{array}$$

Now the attributes of this function.

First, without division or roots, the *domain* is everything: $X = (-\infty, +\infty)$. The second observation is that the values cannot be negative.

Further, let's have a small table of values:

x	$y = f(x) = x^2$
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9

We notice right away that the values (outputs) of f first decrease, up to $x = 0$, and then increase. That can't happen to a linear function!

We also notice something else that distinguishes this function from all linear functions except the constant function:

- Different inputs can produce same output.

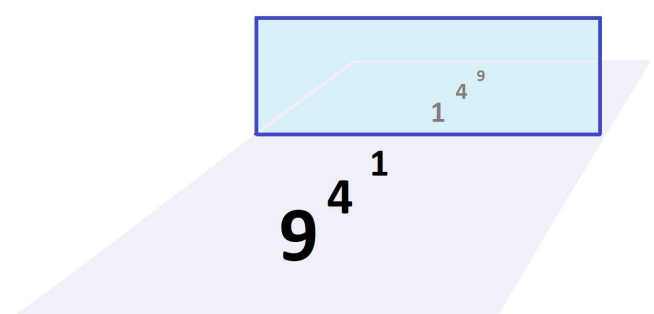
For example:

$$(-1)^2 = 1^2 \,.$$

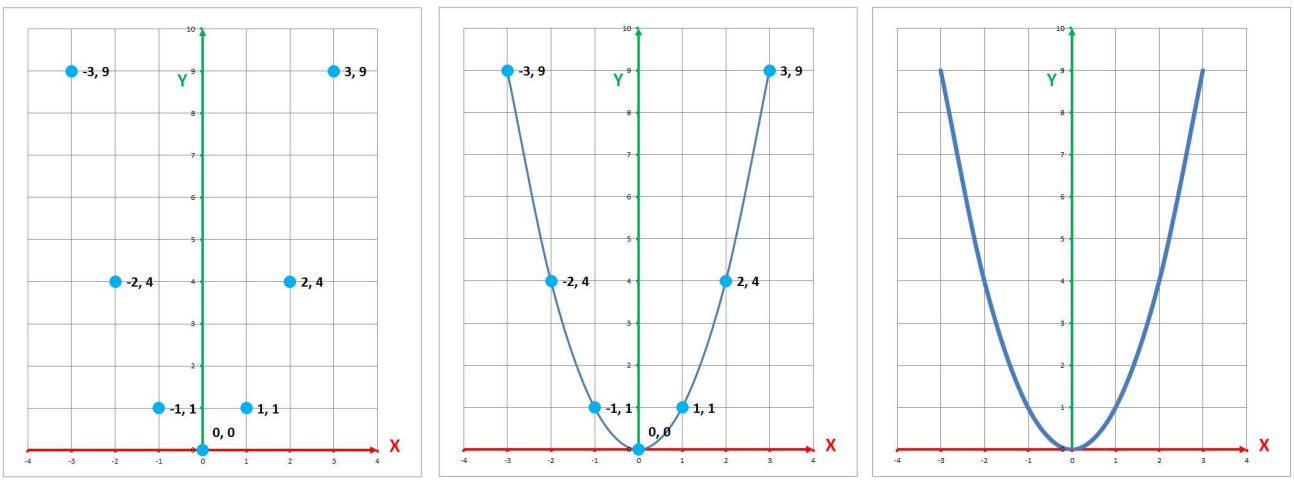
There are many of these. In fact, a pattern starts to emerge:

$$\text{different inputs} \left\{ \begin{array}{c} -3\ 9 \\ \left\{ \begin{array}{c} -2\ 4 \\ \left\{ \begin{array}{c} -1\ 1 \\ 0\ 0 \\ 1\ 1 \end{array} \right\} \\ 2\ 4 \end{array} \right\} \\ 3\ 9 \end{array} \right\} \text{same output!}$$

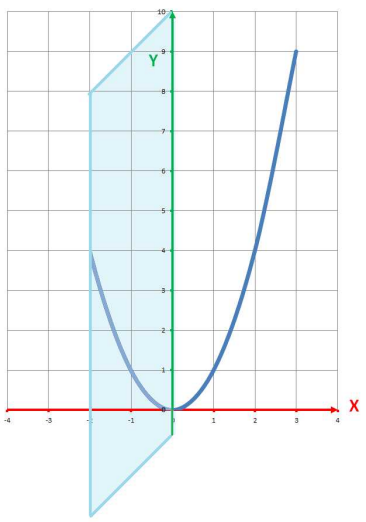
They are paired up! There seems to be a *large-scale symmetry* among the values: They start to repeat themselves – in reverse order – after we pass $x = 0$:



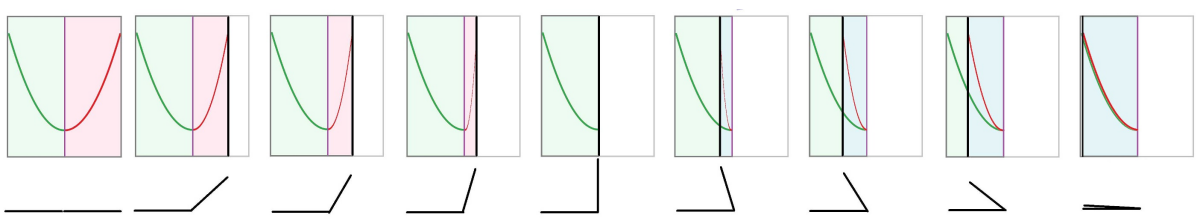
The symmetry becomes vivid once we plot these seven points (left):



When we connect the points to create a curve (middle), we see that its left branch is a *mirror image* of its right branch. The mirror is located on the y-axis:



As an alternative to seeing it in a mirror, we can fold the sheet of paper with the xy-plane on it in half along the y-axis and make one branch of the curve land exactly on top of the other:



Example 2.13.1: transformations as symmetries

Some double-edge swords have this symmetry and some don't:



Exercise 2.13.2

Suggest examples of objects with and without a mirror symmetry.

Exercise 2.13.3

Repeat the above analysis for the following functions:

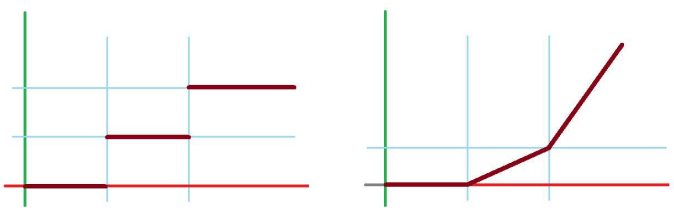
1. $f(x) = x^2 + 1$
2. $g(x) = -x^2$
3. $h(x) = -x^2 + 1$

Example 2.13.4: representing motion by function

Why do we connect these disconnected points by a single curve? Because we try to avoid these two "undesirable" features in the graph:

1. gaps and breaks,
2. corners and cusps.

We do this as follows:

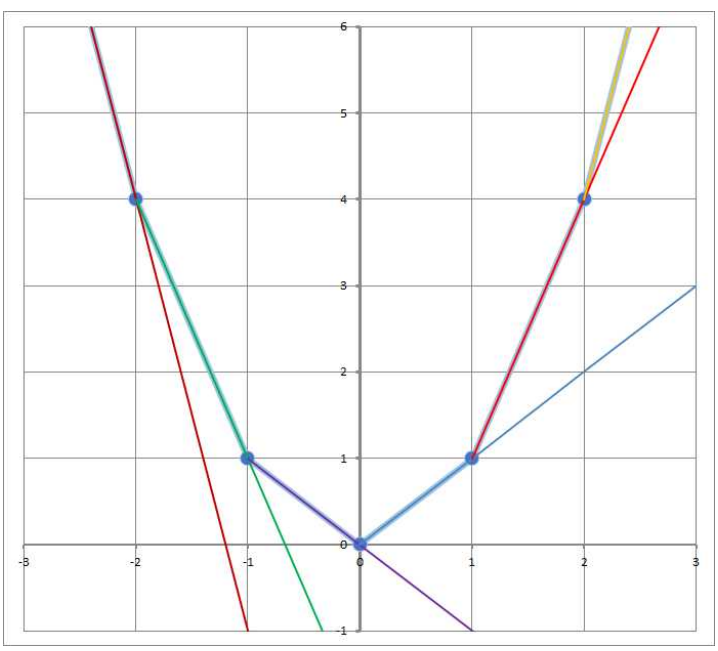


Specifically, if we are to model motion, our function represents the *location as a function of time*. Then these two features represent certain implausible events:

1. The former is an abrupt or even *instantaneous* change of position.
2. The latter is a sudden or even *instantaneous* change of velocity or direction.

It is, however, not unusual for man-made functions to change incrementally. (The two issues are addressed in Chapter 2DC-1 and Chapter 2DC-2, respectively.)

Another feature we notice that distinguishes the square function from a linear function is its slope. It is, in fact, "slopes"; they are different at different locations! Below, we "sample" the function at these 5 points, draw lines through every two consecutive ones, and then compute the slope of each:



These are a couple of computations:

- the slope from $(0,0)$ to $(1,1)$ is $\dfrac{1-0}{1-0} = 1$, but
- the slope from $(1,1)$ to $(2,4)$ is $\dfrac{4-1}{2-1} = 3$.

There is not a single straight line on the graph!

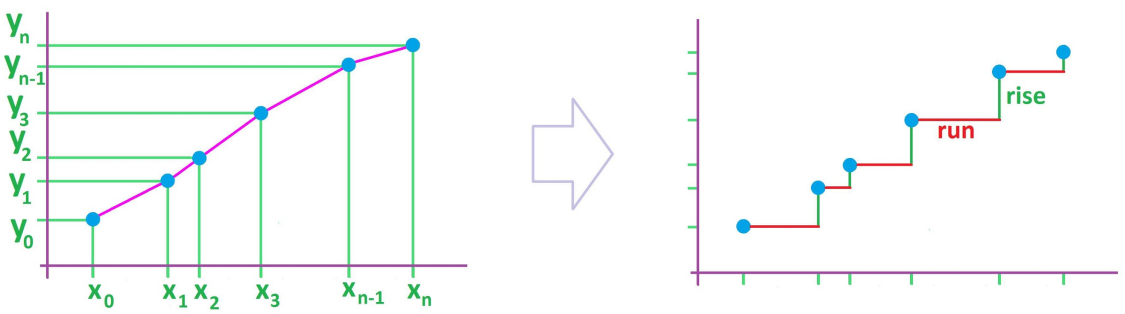
Exercise 2.13.5

Prove this statement.

In general, suppose $y = f(x)$ is a function. Then the slope of f with respect to a pair of points x_0, x_1 is defined to be the slope of the line from $(x_0, f(x_0))$ to $(x_1, f(x_1))$ on the graph of the function, as follows:

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

Here, we are computing slopes for a whole *sequence* of points (x_n, y_n) on the plane. In other words, we are computing the difference quotient introduced in Chapter 1:

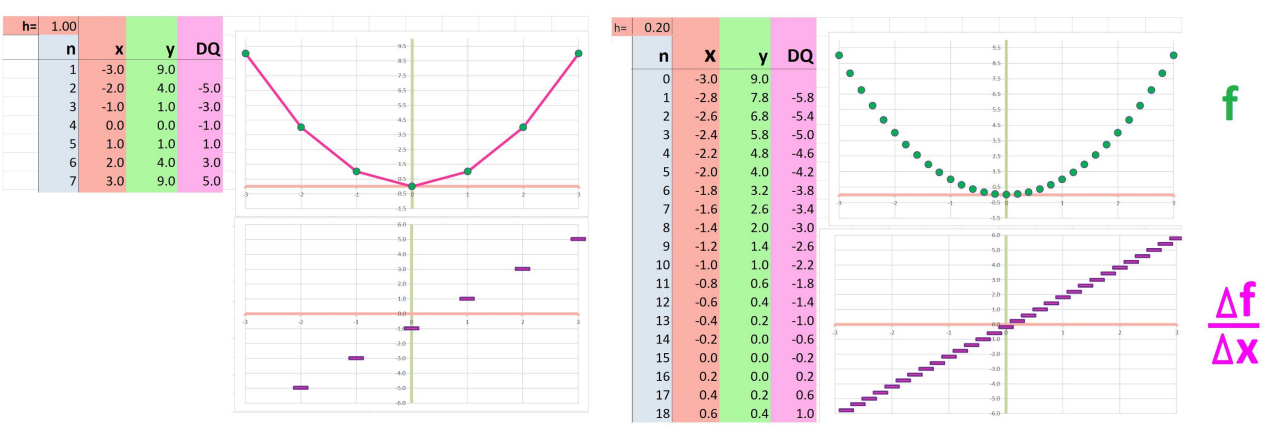


It is the ratio of the sequence of differences of y's and of x's:

$$\frac{\Delta y}{\Delta x} = \frac{y_{n+1} - y_n}{x_{n+1} - x_n}$$

Now these points are points on the graph of a function.

We let the spreadsheet re-do the above computations for $f(x) = x^2$ (left) and plot the slopes in a separate chart below:



We then repeat the procedure for a denser pattern of points (right).

Exercise 2.13.6

What pattern do the slopes exhibit?

Definition 2.13.7: difference quotient of function

Suppose $y = f(x)$ is a function and x_n is a sequence of points within its domain. Then the *difference quotient* of f with respect to this sequence is defined to be

- the sequence of slopes of the lines from $(x_n, f(x_n))$ to $(x_{n+1}, f(x_{n+1}))$ on the graph of the function.

This sequence is denoted and computed as follows:

$$\frac{\Delta f}{\Delta x} = \frac{f(x_{n+1}) - f(x_n)}{x_{n+1} - x_n}$$

The idea of the difference quotient is the idea of the rate of change, and it is central in calculus (Chapter 2DC-2).

With these observations having been made, we know what to look for in every function that comes up.

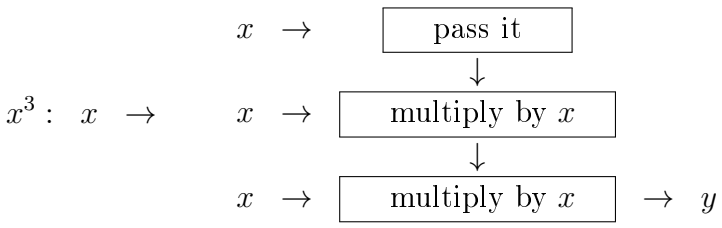
Second, the *cubic function*:

$$\boxed{f(x) = x^3}$$

It is computed via two consecutive multiplications:

$$x^3 = (x \cdot x) \cdot x\,.$$

It is seen in the following flowchart:



While the square function above may be about the *area* of a square $x \times x$, this one may be about the *volume* of a cube $x \times x \times x$.

The *domain* is everything: $X = (-\infty, +\infty)$, again. In contrast, the values *can* be negative:

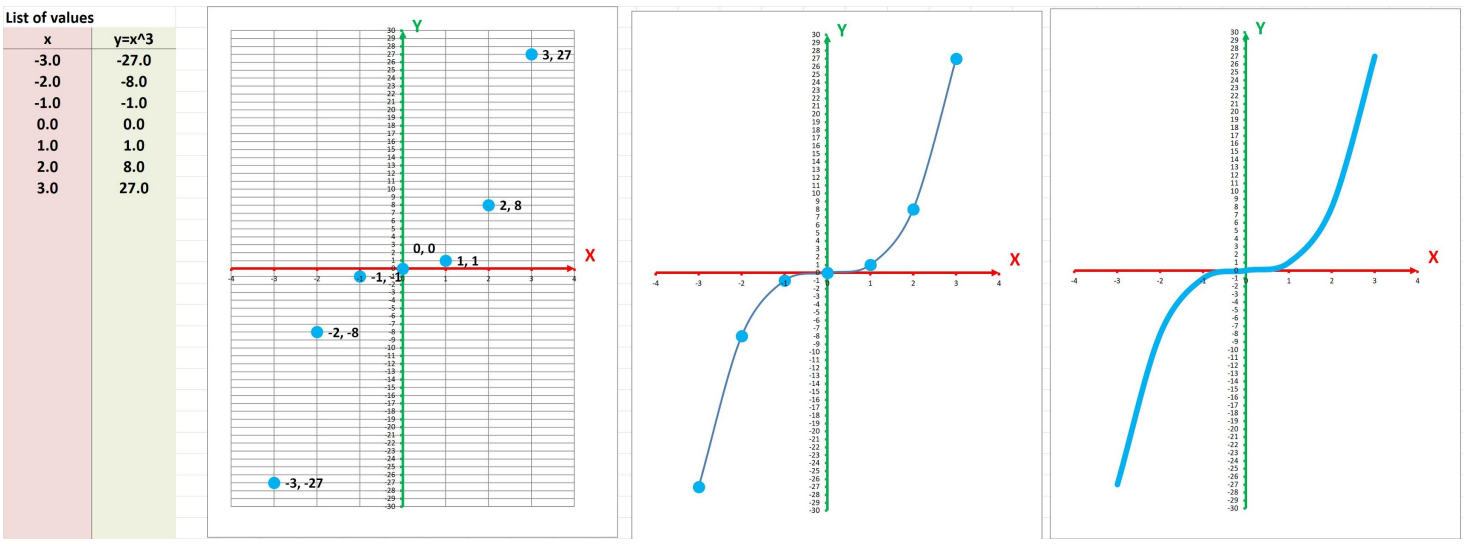
x	$y = f(x) = x^2$
−3	−27
−2	−8
−1	−1
0	0
1	1
2	8
3	27

We notice right away that the values (outputs) of f increase throughout! That's another difference.

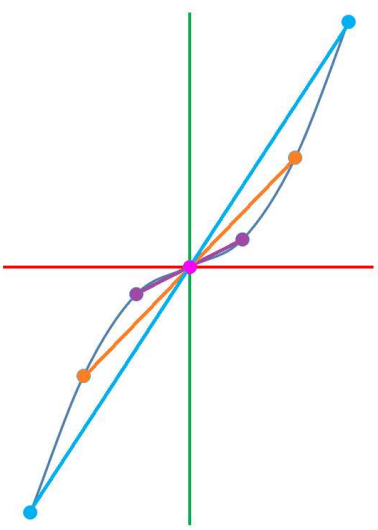
We also notice something else that distinguishes this function from the last: Different inputs *don't* produce same outputs (for example, $(-1)^3 \neq 1^3$), but there is another pattern:

different inputs $\left\{ \left\{ \left\{ \begin{matrix} -3 & -27 \\ -2 & -8 \\ -1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 8 \\ 3 & 27 \end{matrix} \right\} \right\} \right\}$ same output but with opposite sign!

They are paired up as before but differently. There seems to be a large-scale symmetry among the values: They start to repeat themselves – in reverse order with *opposite* signs – after we pass $x = 0$. The symmetry becomes vivid once we plot these seven points:

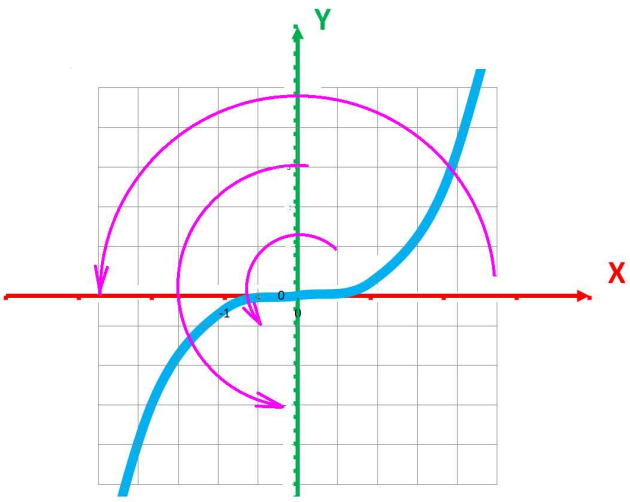


In contrast, the left branch is *not* a mirror image of its right branch! It there another symmetry? Yes, it is called the *central symmetry*:



Here, we draw a line from each point through the origin and then measure the same distance on the other side.

This move can also be executed by a 180-degree rotation around the origin:

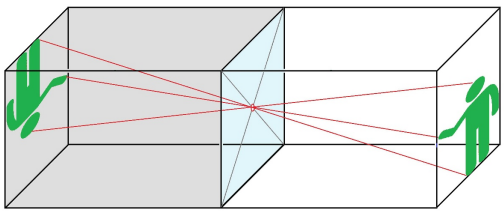


Exercise 2.13.8

Represent this move as two consecutive folds.

Example 2.13.9: transformations as symmetries

The idea of this symmetry can be linked to the "camera obscura":



Exercise 2.13.10

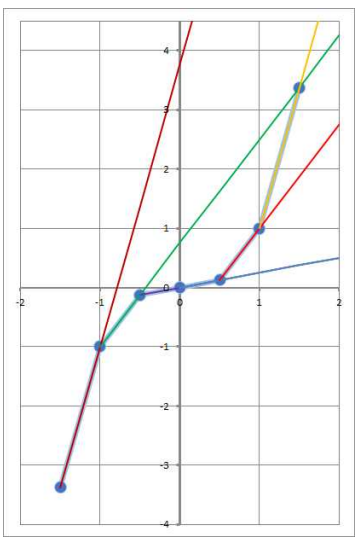
Suggest examples of other objects with and without a central symmetry.

Exercise 2.13.11

Repeat the above analysis for the following functions:

1. $f(x) = x^3 + 1$
2. $g(x) = -x^3$
3. $h(x) = -x^3 + 1$

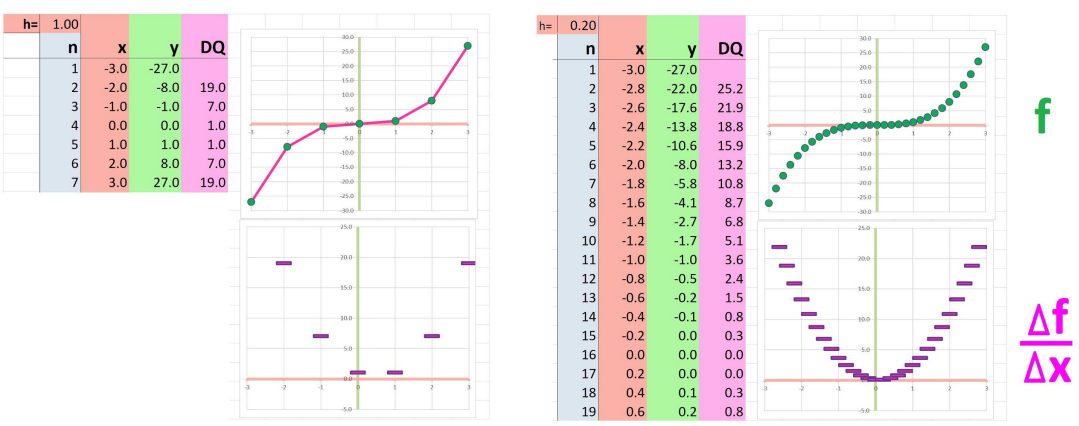
Once again, the *slopes* are different at different locations! Here is a sample of the difference quotient:



Indeed:

- the slope from $(0,0)$ to $(1,1)$ is $\dfrac{1-0}{1-0} = 1$, but
- the slope from $(1,1)$ to $(2,8)$ is $\dfrac{8-1}{2-1} = 7$.

We compute the difference quotient using a spreadsheet for the two values of the increment $h = 1$ and $h = .2$:



h=	1.00			
	n	x	y	DQ
	1	-3.0	-27.0	
	2	-2.0	-8.0	19.0
	3	-1.0	-1.0	7.0
	4	0.0	0.0	1.0
	5	1.0	1.0	1.0
	6	2.0	8.0	7.0
	7	3.0	27.0	19.0

h=	0.20			
	n	x	y	DQ
	1	-3.0	-27.0	
	2	-2.8	-22.0	25.2
	3	-2.6	-17.6	21.9
	4	-2.4	-13.8	18.8
	5	-2.2	-10.6	15.9
	6	-2.0	-8.0	13.2
	7	-1.8	-5.8	10.8
	8	-1.6	-4.1	8.7
	9	-1.4	-2.7	6.8
	10	-1.2	-1.7	5.1
	11	-1.0	-1.0	3.6
	12	-0.8	-0.5	2.4
	13	-0.6	-0.2	1.5
	14	-0.4	-0.1	0.8
	15	-0.2	0.0	0.3
	16	0.0	0.0	0.0
	17	0.2	0.0	0.0
	18	0.4	0.1	0.3
	19	0.6	0.2	0.8

Exercise 2.13.12

What pattern do the slopes exhibit?

As building blocks of more complex functions (Chapter 4), we introduce a whole sequence of functions.

Definition 2.13.13: power functions

The *power functions* are these:

$$\underbrace{x^0 = 1}_{\text{constant}}, \underbrace{x}_{\text{linear}}, \underbrace{x^2}_{\text{quadratic}}, \underbrace{x^3}_{\text{cubic}}, \ldots, \underbrace{x^n}_{n\text{th degree}}, \ldots$$

The power (or the exponent) of x is called the *degree*.

This is how they are computed:

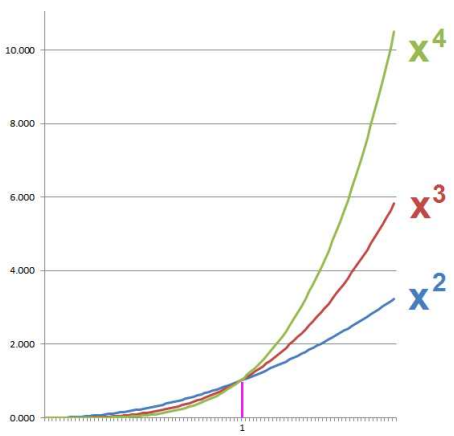
$$x^n: \quad x \to \boxed{\text{pass it}} \downarrow \boxed{\text{multiply by } x} \downarrow \boxed{\text{multiply by } x} \downarrow \cdots \downarrow \boxed{\text{multiply by } x} \to y$$

Warning!

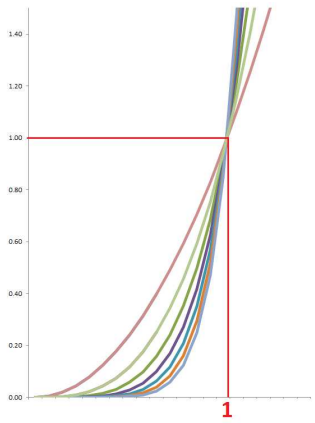
Unlike a geometric progression, the multiplication in a power function is repeated a *fixed* number of times.

The same questions are asked and answered about these new functions: The domains are all real numbers, etc.

The magnitude of the degree affects the shape of the graph:



They all meet at $(0, 0)$ and $(1, 1)$. Also, the higher the degree, the slower the graph grows from $x = 0$ and the faster it rises from $x = 1$:



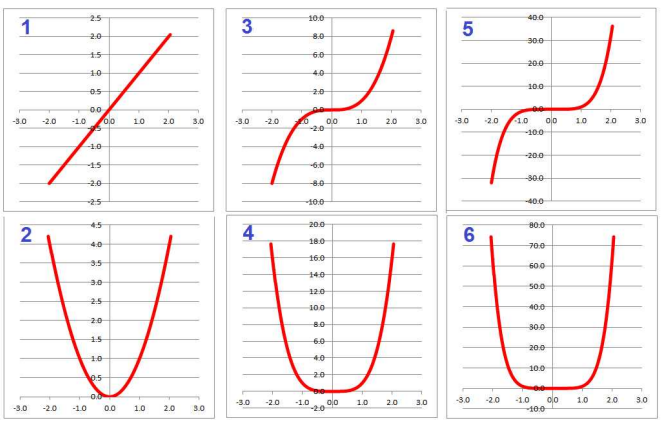
Exercise 2.13.14

If we cross all these graphs with the line $x = r$, what do these points of intersection form?

Exercise 2.13.15

What does the graph of $y = x^{100}$ look like between $x = 0$ and $x = 1$?

Here is another important pattern:



One might be tempted to say that all the graphs in the second row – these are the even powers – look alike, i.e., are parabolas.

Warning!

The graph of $y = x^2$ is a parabola, but the graphs of $y = x^4$ and $y = x^6$ are not.

We recognize that x^4, x^6, ... have flatter bottoms (relative to the growth that follows). When the power is odd, the graphs "look like" that of x^3 but also flatter around 0. Thus, the "evenness" of the degree significantly affects the shape of the graph:

- The graphs of the even degree powers have mirror symmetry about the y-axis.
- The graphs of the odd degree powers have central symmetry about the origin.

Understandably, functions with the former kind of symmetry are called "odd" and the latter "even". These functions are discussed in Chapter 4.

This is the summary of the analysis:

$y = x^n$	domain	values	trend	long-term trend	symmetry
odd powers	$\mathbf{R}$	$-+$	down-up	infinity	central
even powers	$\mathbf{R}$	$++$	up-up	infinity	mirror

Exercise 2.13.16

Sketch what you think $y = x^7$ and $y = x^8$ look like.

No division until now!

Next, the *reciprocal function*:

$$\boxed{f(x) = \frac{1}{x}}$$

We see the difference right away. We can't divide by 0 and, therefore, the *domain* is

$$X = \{x : \ x \neq 0\}.$$

Because of the hole in the domain, the graph consists of *two branches*.

A few values are listed below:

x	$y = f(x) = 1/x$
-3	$-1/3$
-2	$-1/2$
-1	-1
0	NOTHING!
1	1
2	1/2
3	1/3

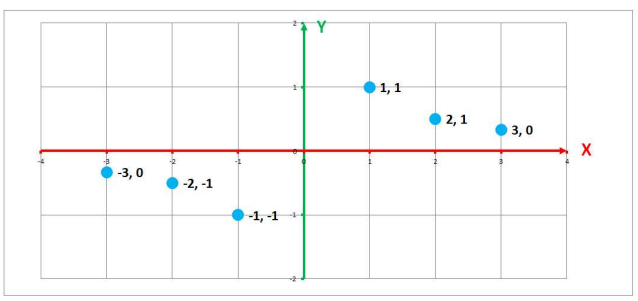
We notice right away that the values (outputs) of f decrease throughout either of the two halves of the domain!

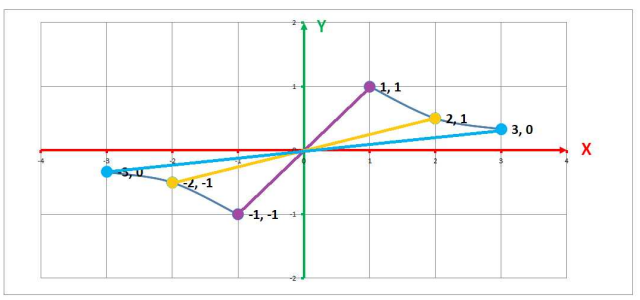
Just as with x^3, different inputs *don't* produce same outputs but there is another pattern:

$$\text{different inputs} \left\{ \left\{ \left\{ \begin{array}{cc} -3 & -1/3 \\ -2 & -1/2 \\ -1 & -1 \end{array} \right\} \right\} \right\} \text{outputs with opposite signs!}$$

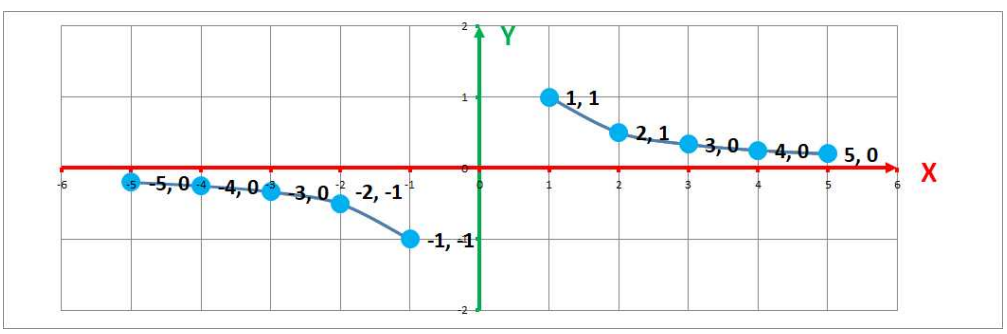
$$\begin{array}{cc} 1 & 1 \\ 2 & 1/2 \\ 3 & 1/3 \end{array}$$

It's the central symmetry again:





Adding a couple of points at either end of the graph shows that the graph starts to approach the x-axis, seems to almost merge with it, but never actually reaches it:

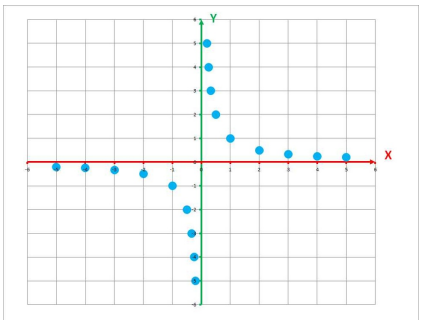


We say that $y = 0$ (the x-axis) is a "horizontal asymptote" of the graph (Chapter 2DC-2).

But what is going on closer to the *hole in the domain*, 0? Let's *insert* points in the middle:

x	-3	-2	-1	$-1/2$	$-1/3$	$-1/4$	$-1/5$	$\circ$	$1/5$	$1/4$	$1/3$	$1/2$	1	2	3
$y = 1/x$	$-1/3$	$-1/2$	-1	-2	-3	-4	-5	|	5	4	3	2	1	$1/2$	$1/3$

The result shows that the graph starts to approach the y-axis, seems to almost merge with it, but never actually reaches it:



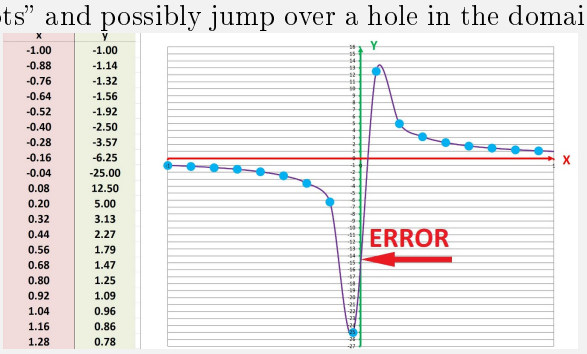
We say that $x = 0$ (the y-axis) is a "vertical asymptote" of the graph (Chapter 2DC-2). This curve is called a *hyperbola*.

Exercise 2.13.17

Prove that the axes are never touched by the graph.

Warning!

Relying on a computer to plot such a graph might cause errors as it might attempt to "connect the dots" and possibly jump over a hole in the domain:



This is the difference quotient $\dfrac{\Delta f}{\Delta x}$, i.e., the sampled slopes, of $1/x$ (bottom row):



Exercise 2.13.18

What does the other half look like?

Exercise 2.13.19

The graph has also a mirror symmetry; point it out.

Exercise 2.13.20

Take another look at the last exercise.

As there are more and more functions, we can't devote as much time and attention to each and every one of them. Often, we will get only a bird's eye view of *classes* of functions.

In light of Chapter 1, in addition to the positive power functions, we now have the *negative power functions*, as the reciprocals of the power functions:

$$x^{-1} = \frac{1}{x^1},\ x^{-2} = \frac{1}{x^2},\ x^{-3} = \frac{1}{x^3},\ ...,\ x^{-n} = \frac{1}{x^n},\ ...$$

They are easy to compute if we have the former functions available. For example, this is how one computes $1/x^3$:

$$x \ \rightarrow\ \boxed{\text{cube it}}\ \rightarrow\ \boxed{\text{take its reciprocal}}\ \rightarrow\ y$$

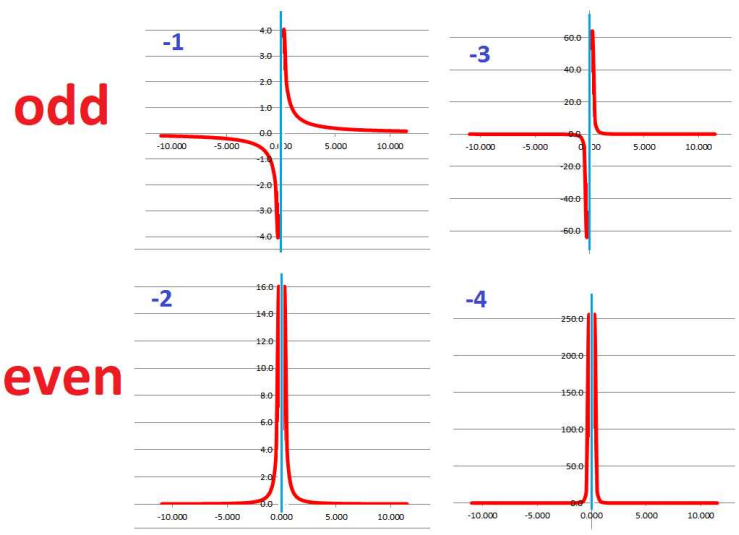
Exercise 2.13.21

What happens if we change the order of operations?

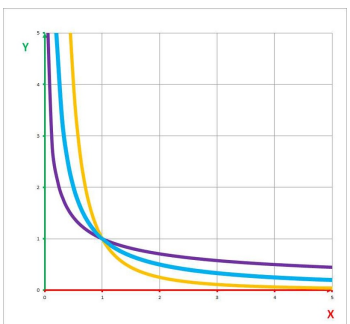
The main difference from the positive powers is in the domains as 0 must be excluded. The *domains* of all of these functions are the same:

$$X = \{x :\ x \neq 0\}\,.$$

Instead of just recognizing patterns in a behavior of a single function, we try to see them in a whole *group of functions*. For example, this is how the magnitude of the degree affects the shape of the graph:



The higher the degree, the faster the graph drops from $x = 0$ and the slower it declines from $x = 1$. They all meet at $(1, 1)$:



We recognize that as we move along the sequence of functions $1/x^3$, $1/x^4$, ..., the graphs are getting *flatter and flatter*, almost horizontal! When the power is odd, the graph "looks like" that of $1/x$ but is located closer to the x-axis and farther from the y-axis. Once again, the evenness of the degree significantly affects the shape of the graph:

- The graphs of the even degree powers have mirror symmetry about the y-axis.
- The graphs of the odd degree powers have central symmetry about the origin.

This is the summary of the analysis:

$y = x^{-n}$	domain	values	trend	long-term trend	symmetry
odd powers	$x \neq 0$	$-+$	down, down	0	central
even powers	$x \neq 0$	$++$	up, down	0	mirror

Warning!

We can't simply say that the values of an odd degree negative power function decrease, not across the hole in the domain.

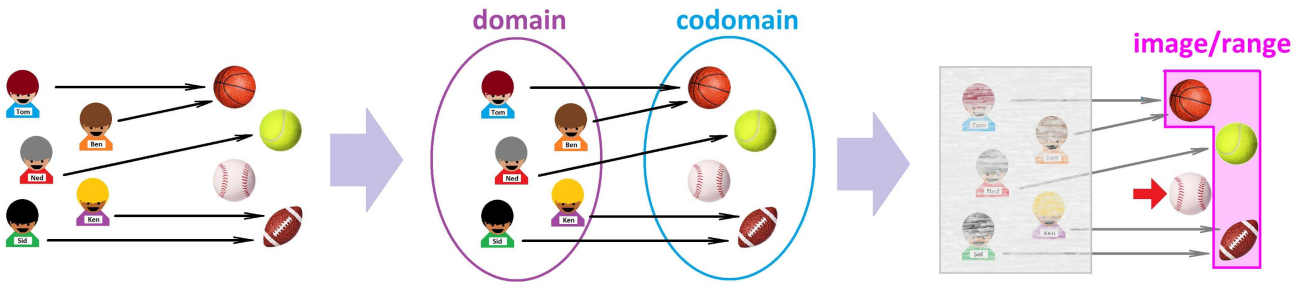
Exercise 2.13.22

Sketch what you think $y = 1/x^5$ and $y = 1/x^6$ look like.

These properties and features of functions are so important that we will continue to look for them in every function that comes up. Conversely, the simple functions presented in this section are the go-to examples of the features we have discovered.

2.14. The image: the range of values of a function

Let's go back to the set X of boys, the set of balls Y, and the "I prefer" function F from X to Y. A simple question we may ask about it is: What do the boys like *as a group*? It has a simple answer, a list: basketball, tennis, and football. We just have to look at the arrow and record each element of Y that has an arrow drawn toward it:



This set is a subset of the codomain Y:

$$V = \{ \text{ basketball, tennis, football } \} \subset Y\,.$$

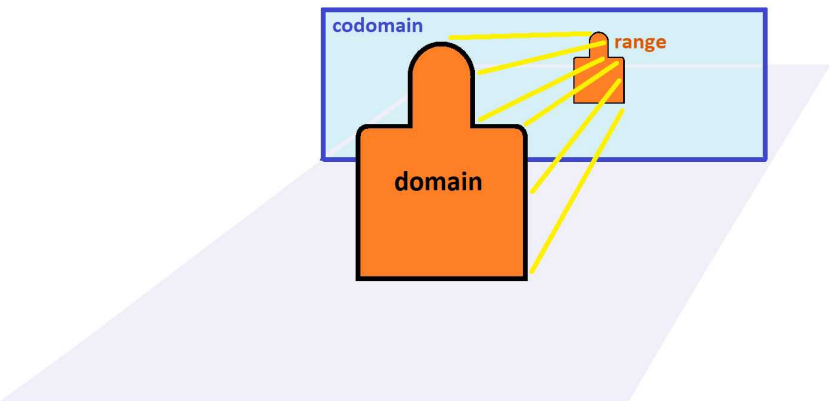
While the latter is the set of all possible or potential values of F, the former is the set of *actual* values. In other words, this is the range of values of the function. It can be, but is not in this case, the whole codomain.

Definition 2.14.1: image/range of function

The *image*, or the *range*, of a function $F : X \to Y$ is the set of all of its values, i.e.,

$$\{y : \ F(x) = y\,, \quad \texttt{FOR SOME } x\}\,.$$

We see the image of the domain reflected in the codomain:

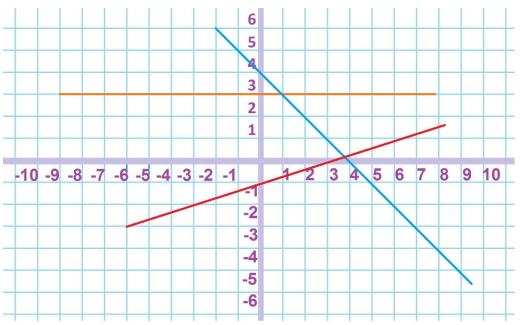


In the definition, we test each y: Is there a corresponding x? If there is, we add this y to the set.

Note that if we keep the values but change the original codomain of F to its range V, we have a new function $G : X \to V$.

Now numerical functions. You get an idea about the range by simply looking at the y-column of the *table of values* of the function (just as looking at the x-column gives you an idea about the domain). However, finding the set explicitly requires some algebra.

Linear functions are simple:



We need to try, if possible, to find an x for each y. The computations are easy too:

$$y = mx + b \implies x = \frac{y - b}{m}\,,$$

for $m \neq 0$. So, there is an x for every y! We have proven the former part of the following:

Theorem 2.14.2: Range of Linear Function

The range of a linear function $y = mx + b$ is the set of real numbers, $V = \mathbf{R}$, when $m \neq 0$; otherwise, the range is a single point, $V = \{b\}$.

Exercise 2.14.3

Prove the latter part.

Exercise 2.14.4

State the theorem as an equivalence (an "if-and-only-if" statement).

Example 2.14.5: range of x^2 and x^3

Can we make the same argument for $f(x) = x^2$? Of course not: Squares can't be negative! But what about the rest of y's? We attempt to solve the equation, for each y:

$$y = x^2 \implies x = \sqrt{y} \quad \texttt{AND} \quad y \geq 0\,.$$

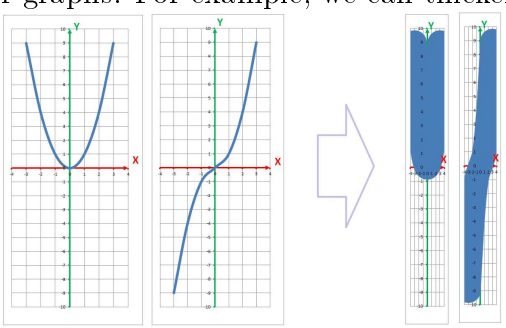
Therefore, the range of x^2 is

$$\{y:\ y \geq 0\} = [0, +\infty)\,.$$

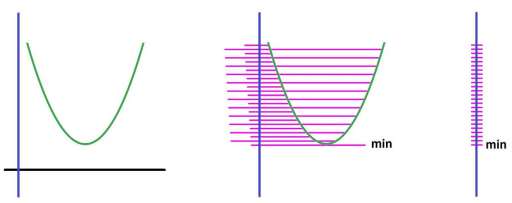
What about x^3? It works for *any* y:

$$y = x^3 \implies x = \sqrt[3]{y}\,.$$

Why such a difference? In addition to the algebra above, we will appreciate the difference between the two functions by examining their graphs. For example, we can thicken them and shrink the xy-plane:



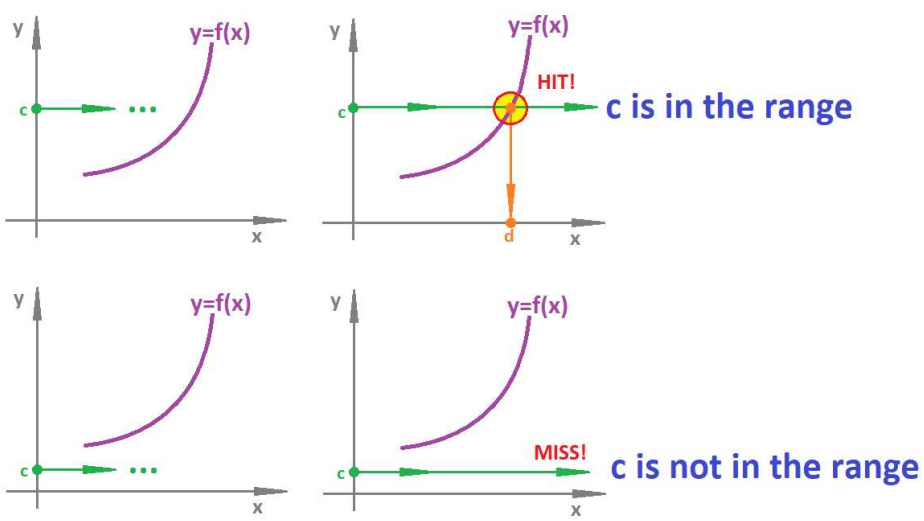
So, the vertical "spread" of the graph gives us the range (the horizontal "spread" of the graph gives us the domain). This is how the range of $y = x^2$ is seen as a ray in the y-axis:



Exercise 2.14.6

Use these two methods to find the domain of these functions.

Generally, to find the range of a numerical function, the graph of which is supplied, we test y's one at a time. From each of them, we draw a horizontal line and note whether it crosses the graph:

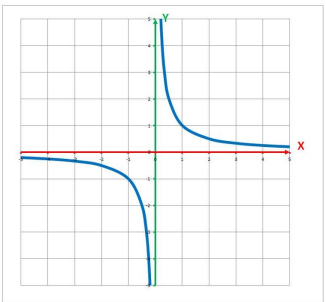


Example 2.14.7: range of reciprocal

The example of $y = 1/x$ is a bit more complex. What y can come from this formula? To answer, find x:

$$y = 1/x \implies x = 1/y \quad \texttt{AND} \quad y \neq 0\,.$$

In other words, the equation $1/x = 0$ has no solution. We come to the same conclusion by examining the graph and discovering that it cannot touch the x-axis:



Therefore, the range is

$$V = \{y:\ y \neq 0\}\,.$$

By the same method we discover that the range is exactly the same for all reciprocals of odd powers.

Exercise 2.14.8

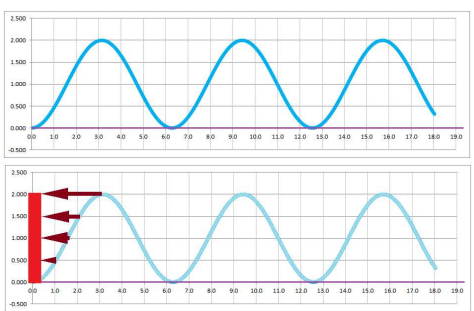
Find the range for all reciprocals of even powers.

So, these graphs cannot touch or cross the x-axis, and that is the same as to say that 0 isn't in the range. Likewise, their graphs cannot touch or cross the y-axis and that is the same as to say that 0 isn't in the domain. That's the analogy – and the symmetry – of the problems of the domain and the range (not the codomain). It is the symmetry between the x-axis and the y-axis in the xy-plane.

Warning!

Range $\neq$ codomain.

Any set (in the y-axis) can be the range of some function. For example, this function's range is a closed bounded interval:

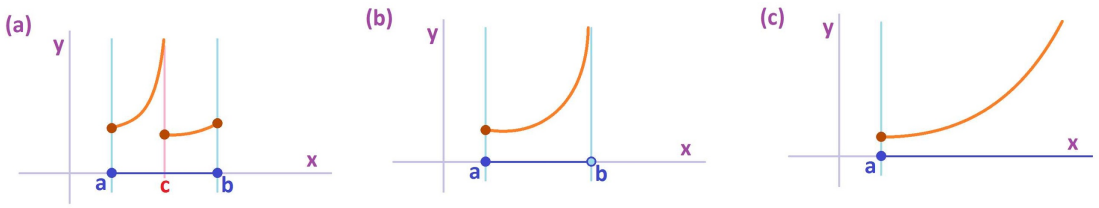


The following will be routinely used.

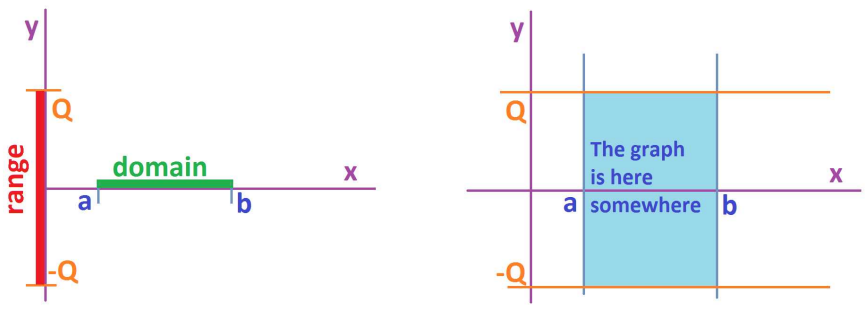
Definition 2.14.9: bounded and unbounded functions

If the range of a function is a bounded set, the function is called *bounded*; otherwise, it is *unbounded*.

The linear polynomials are unbounded, and so are all quadratic polynomials. These are some ways a function can exhibit unbounded behavior:

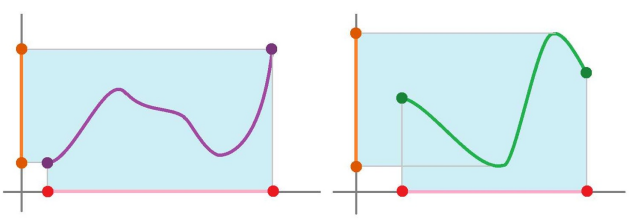


If the domain and the range are intervals, the graph of the function is contained in the rectangle with these sides:

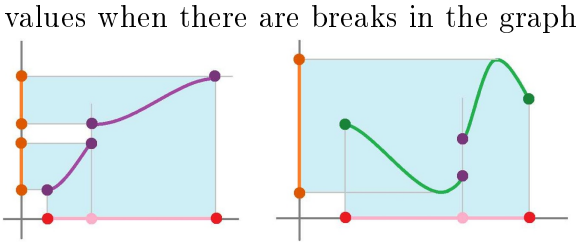


Example 2.14.10: range from graph

It is often the case that the domain is an interval (the codomain is typically chosen to be $Y = \mathbf{R}$). And so is the range:

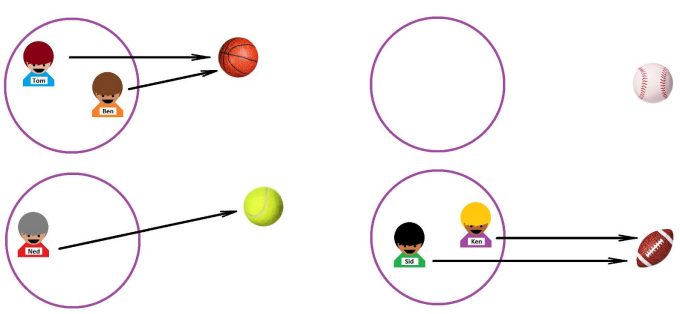


However, the range may skip values when there are breaks in the graph:



(This issue is discussed in Chapter 2DC-1.)

Another question we can ask about the boys and the balls is: Who likes basketball? or baseball, etc.? We just look at the arrow, or arrows, that is drawn towards this ball and note where it comes from. The result is a subset of X. This is what happens with the above example:



Definition 2.14.11: preimage of value

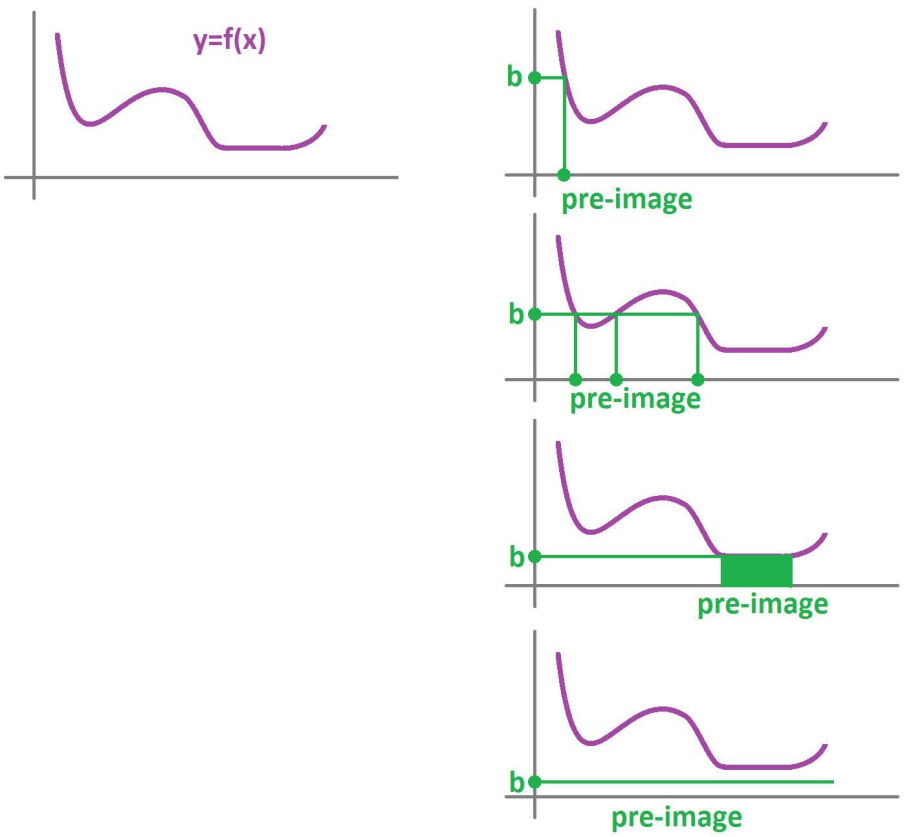
The *preimage* of an element b in a set Y under a function $F : X \to Y$ is the set of all x's the value of which under F is b, i.e.,

$$\{x : \ F(x) = b\}\,.$$

In other words, we are *solving equations* again.

We carry out this computation for every ball. We discover, in particular, that the preimage of baseball is the empty set. This is always the case with outputs outside the range!

The picture below illustrates how to find the preimage of a point of a numerical function:



Some answers we already know:

- The preimages under a constant function are empty with an exception of a single value, the preimage of which is the whole x-axis.
- The preimages under linear (non-constant) polynomials are single points.
- The preimages under even powers are two-point sets for positive y's, a single point for $y = 0$, and empty for negative y's.
- The preimages under odd powers are single points.

Exercise 2.14.12

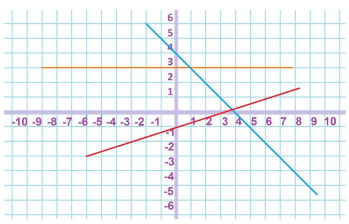
Prove the above statements.

Exercise 2.14.13

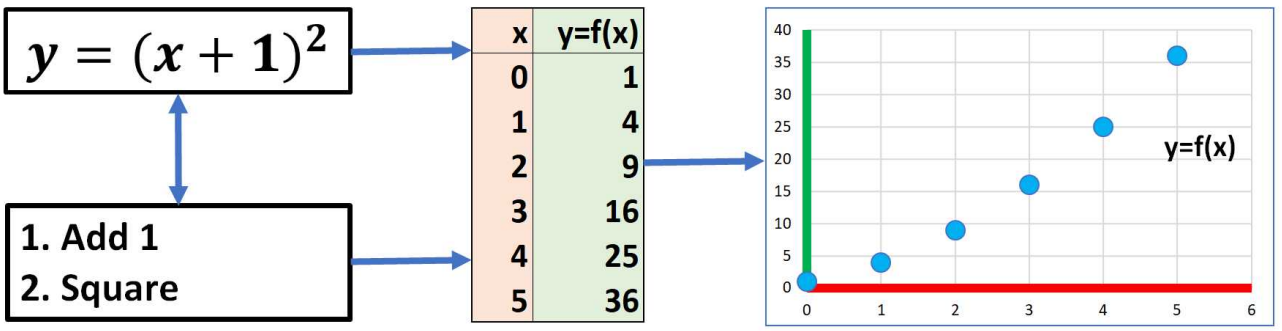
What are the preimages of the reciprocals of the powers?

Exercise 2.14.14

The segments of straight lines below are graphs of three functions. Find the domains and the ranges of these functions:



Below is a summary of our treatment of functions in this chapter:



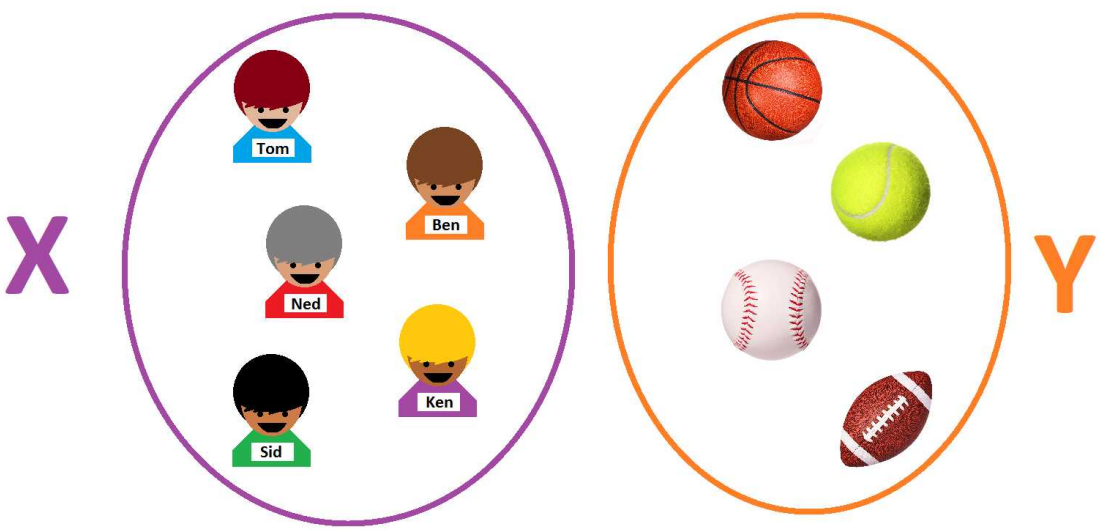
Chapter 3: Compositions of functions

Contents

3.1 Operations on sets 199
3.2 Piecewise-defined functions 207
3.3 Numerical functions are transformations of the line 217
3.4 Functions with regularities: one-to-one and onto 224
3.5 Compositions of functions 234
3.6 The inverse of a function 244
3.7 Units conversions and changes of variables 259
3.8 Transforming the axes transforms the plane 264
3.9 Changing a variable transforms the graph of a function 273
3.10 The graph of a quadratic polynomial is a parabola 284

3.1. Operations on sets

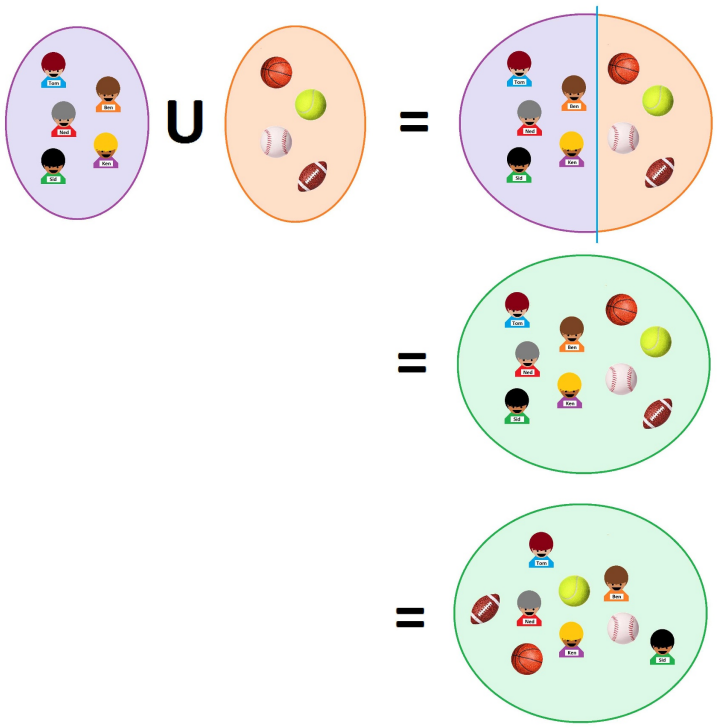
Let's go back to our example of the five boys that form a set and another set is the set of these four balls:



They are just lists without repetitions:

$$\begin{aligned} X &= \{ \text{Tom} , \text{Ken} , \text{Sid} , \text{Ned} , \text{Ben} \}; \text{ and} \\ Y &= \{ \text{basketball} , \text{tennis} , \text{baseball} , \text{football} \}. \end{aligned}$$

We can form a new set that contains all the elements of the two sets, as follows:



We can merge the two sets together and, once the separator is gone, move the elements around. The following concept will be routinely used.

Definition 3.1.1: union of sets

The *union* of any two sets X and Y is the set that consists of the elements that belong to either X or Y. It is denoted by

$$X \cup Y$$

The most common case is when both sets are *subsets* of the same set. For example, all of these are subsets of $X \cup Y$:

$$\begin{array}{ll} T = \{\text{ Tom }\}, \quad A = \{\text{ Tom, Ken }\}, \quad Q = \{\text{ Tom, Ken, Sid }\}, \quad \ldots & \subset X \\ B = \{\text{ basketball }\}, \quad V = \{\text{ basketball, tennis }\}, \quad U = \{\text{ basketball, tennis, baseball }\}, \quad \ldots & \subset Y \\ \Longrightarrow & \\ \{\text{ Tom }\} \cup \{\text{ basketball }\} = \{\text{ Tom, basketball }\} & \subset X \cup Y \\ \{\text{ Tom, Ken }\} \cup \{\text{ basketball, tennis }\} = \{\text{ Tom, Ken, basketball, tennis }\} & \subset X \cup Y \end{array}$$

Exercise 3.1.2

Present the unions of all *pairs* of sets shown above. Now, define the unions of *triples* of sets and present them. Continue.

So, we check for each x the following:

$$x \text{ belongs to } X \quad \texttt{OR} \quad x \text{ belongs to } Y$$

Warning!

"X `OR` Y" means "X `OR` Y or both".

Exercise 3.1.3

Solve for Y:

$$X \cup Y = X\,.$$

So, finding the union of two sets given as lists is very simple:

- ▶ We merge the lists removing repetitions.

These repetitions comprise the "overlap" of the two sets:

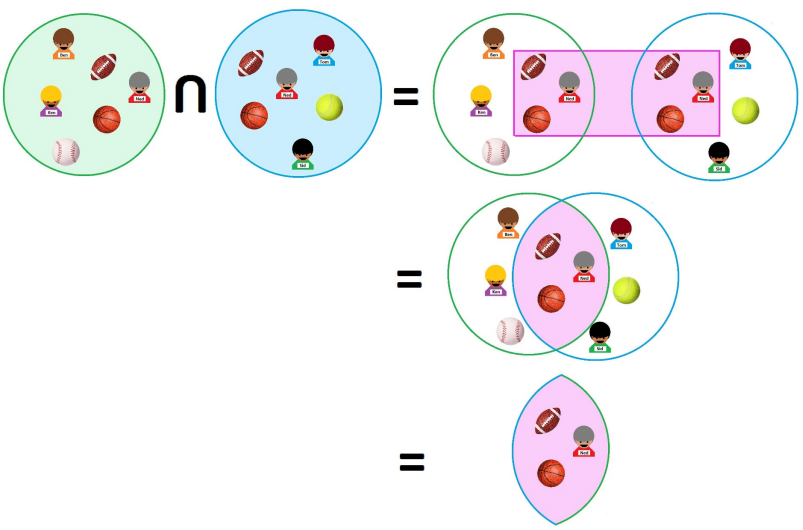
$$\begin{aligned}\{\text{ tennis, Tom, Ken }\} \cup \{\text{ basketball, tennis, Tom }\} &= \{\text{ tennis, Tom, Ken, basketball, tennis, Tom }\}\\ &= \{\text{Tom, Ken, basketball, tennis }\}.\end{aligned}$$

Example 3.1.4: unions of lists

We find the union in two steps:

$$\begin{aligned}\{1,2,3,4\} \cup \{3,4,5,6,7\} &= \{1,2,3,4,\quad 3,4,5,6,7\}, && \text{merge the lists...}\\ &= \{1,2,3,4,5,6,7\}, && \text{then remove repetitions.}\end{aligned}$$

Next, the "overlap" itself. We look at what the two sets have *in common*:



We find the repeated parts and merge each into one; the rest is thrown out. The following concept will be routinely used.

Definition 3.1.5: intersection of sets

The *intersection* of any two sets X and Y is the set that consists of all the elements that belong to both X and Y. It is denoted by

$$X \cap Y$$

Here is an example of such a computation:

$$\{\text{ tennis, Tom, Ken }\} \cap \{\text{ basketball, tennis, Tom }\} = \{\text{ tennis, Tom }\}.$$

So, we check for each x the following:

$$x \text{ belongs to } X \quad \texttt{AND} \quad x \text{ belongs to } Y$$

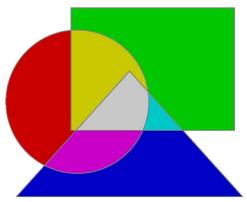
Exercise 3.1.6

Solve for Y:

$$X \cap Y = X\,.$$

Example 3.1.7: colors

The *primary colors* are red, green, and blue. The rest of the colors are seen as "combinations" of these three:



We can also interpret them as intersections.

Recall that the set-building notation is used to create a set by stating a condition these numbers are supposed to satisfy:

$$Z = \{x : \quad \text{condition for } x\ \}\,.$$

We consider what happens to sets given this way under these two operations.

Example 3.1.8: solution sets of systems of equations

Numerical sets are subsets of the real line $\mathbf{R}$ and some of them came from solving these equations, as their *solution sets*:

equation	solution set	simplified
$x^2 - 3x + 2 = 0$	$X = \{x : \ x^2 - 3x + 2 = 0\}$	$= \{1, 2\}$
$x^2 = 1$	$Y = \{x : \ x^2 = 1\}$	$= \{-1, 1\}$

What if we have both equations to be satisfied *at the same time*? We are interested in the set:

$$\begin{aligned} Z &= \{x : \ x^2 - 3x + 2 = 0 \quad \texttt{AND} \quad x^2 = 1\ \} \\ &\qquad\qquad \Longrightarrow \qquad x \text{ belongs to both } X \ \texttt{AND}\ Y\,! \\ Z &= X \ \cap \ Y \\ &= \{x : \ x^2 - 3x + 2 = 0\} \ \cap \ \{x : \ x^2 = 1\} \\ &= \{1, 2\} \ \cap \ \{-1, 1\} \\ &= \{1\}\,. \end{aligned}$$

Now, what if we need *just one* of the equations to be satisfied? We are interested in the set:

$$\begin{aligned} Z &= \{x : \ x^2 - 3x + 2 = 0 \quad \texttt{OR} \quad x^2 = 1\ \} \\ &\qquad\qquad \Longrightarrow \qquad x \text{ belongs to } X \ \texttt{OR}\ Y\,! \\ Z &= X \ \cup \ Y \\ &= \{x : \ x^2 - 3x + 2 = 0\} \ \cup \ \{x : \ x^2 = 1\} \\ &= \{1, 2\} \ \cup \ \{-1, 1\} \\ &= \{-1, 1, 2\}\,. \end{aligned}$$

So, sometimes the condition that defines a set splits into *two* conditions:

$$Z = \{x : \quad \text{satisfies first condition} \quad \texttt{AND} \quad \text{satisfies second condition}\ \}\,.$$

The word "AND" is capitalized in order to emphasize that the set contains only those x's that satisfy *both* conditions simultaneously. Then we can see also that there are *two* sets:

$$X = \{x : \quad \text{satisfies first condition}\ \} \text{ and } Y = \{x : \quad \text{satisfies second condition}\ \}\,,$$

one for either condition. We are interested in their intersection:

$$Z = X \cap Y = \{x: \quad \text{satisfies first condition}\} \cap \{x: \quad \text{satisfies second condition}\}.$$

Example 3.1.9: solution sets of inequalities

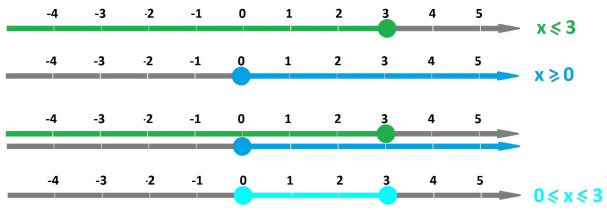
These conditions can also be inequalities. Let's consider this "double inequality":

$$0 \leq x \leq 3\,.$$

Its solution set is all x's that satisfy the inequality. However, there are *two* inequalities in reality and two solution sets:

$$\begin{array}{lll} & \{x:\ 0 \leq x \leq 3\} & \\ = & \{x:\ x \geq 0 & \texttt{AND} \quad x \leq 3\} \\ = & \{x:\ x \geq 0\} & \cap \quad \{x:\ x \leq 3\} \\ = & [0, +\infty) & \cap \quad (-\infty, 3] \\ = & & [0, 3]\,. \end{array}$$

The geometric interpretation is shown below:

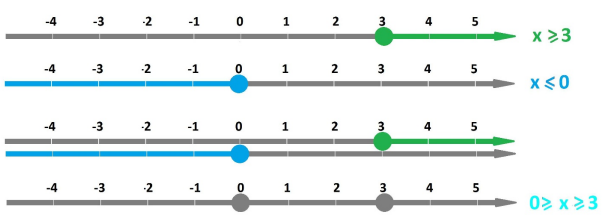


What if we "flip" both signs of this inequality:

$$0 \geq x \geq 3\ ?$$

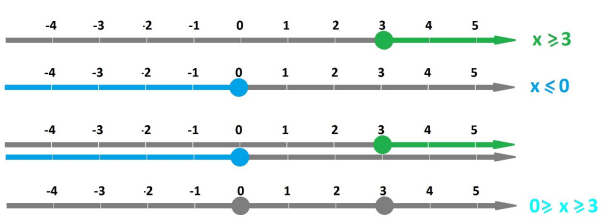
Then:

$$\begin{array}{lll} & \{x:\ 0 \geq x \geq 3\} & \\ = & \{x:\ x \leq 0 & \texttt{AND} \quad x \geq 3\} \\ = & \{x:\ x \leq 0\} & \cap \quad \{x:\ x \geq 3\} \\ = & (-\infty, 0] & \cap \quad [3, +\infty) \\ = & & \varnothing\,. \end{array}$$



What if we are interested in when the original inequality is *not* satisfied? This is our set:

$$\begin{array}{lll} & \{x:\ 0 \leq x \leq 3, \quad \texttt{FALSE}\} & \\ = & \{x:\ x \geq 0 & \texttt{AND} \quad x \leq 3, \quad \texttt{FALSE}\} \\ = & \{x:\ x \geq 0, \quad \texttt{FALSE} & \texttt{OR} \quad x \leq 3, \quad \texttt{FALSE}\} \\ = & \{x:\ x \leq 0 & \texttt{OR} \quad x \geq 3\} \\ = & \{x:\ x \leq 0\} & \cup \quad \{x:\ x \geq 3\} \\ = & (-\infty, 0] & \cup \quad [3, +\infty)\,. \end{array}$$



Exercise 3.1.10

What happens if you flip only one of the two inequalities above?

As a summary, let's rephrase our two definitions.

Definition 3.1.11: union and intersection of sets

$$X \cup Y = \{x : \ x \text{ belongs to } X \quad \texttt{OR} \quad x \text{ belongs to } Y\}$$
$$X \cap Y = \{x : \ x \text{ belongs to } X \quad \texttt{AND} \quad x \text{ belongs to } Y\}$$

Warning!

The word `OR` in mathematics is always meant *inclusively*: not as in "black or white" but as in "cream or sugar".

When the two sets are defined via conditions, we have the following for their intersection:

$$\begin{aligned} & \{x : \ \text{first condition} \quad \texttt{AND} \quad \text{second condition } \} \\ = \ & \{x : \ \text{first condition } \} \cap \{x : \ \text{second condition } \}; \end{aligned}$$

and for the union:

$$\begin{aligned} & \{x : \ \text{first condition} \quad \texttt{OR} \quad \text{second condition } \} \\ = \ & \{x : \ \text{first condition } \} \cup \{x : \ \text{second condition } \}. \end{aligned}$$

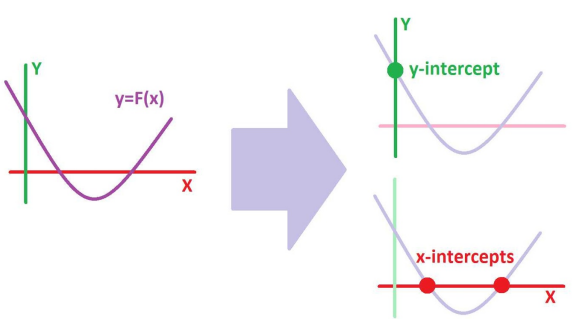
We next consider *subsets of the plane*.

Example 3.1.12: intercepts

For a numerical function F, its *graph* is the following set presented via the set-building notation:

$$\{(x, y) : \ y = F(x)\} \subset \mathbf{R}^2 .$$

Now, the two especially important subsets of the xy-plane are its two *axes*. Now, let's put these together:



The *x-intercepts* of F are the elements of the intersection of the graph of F with the x-axis:

$$\{ \ x\text{-intercepts } \} = \text{graph of } F \ \cap \ x\text{-axis}.$$

Meanwhile, the *y-intercept* of F is the (possible) element of the intersection of the graph of F with the y-axis:

$$\{ \ y\text{-intercept } \} = \text{graph of } F \ \cap \ y\text{-axis}.$$

Warning!

Don't confuse an element of a set and a one-element set.

Exercise 3.1.13

Give examples of functions for which these sets are empty.

Exercise 3.1.14

Explain the meaning of these sets:

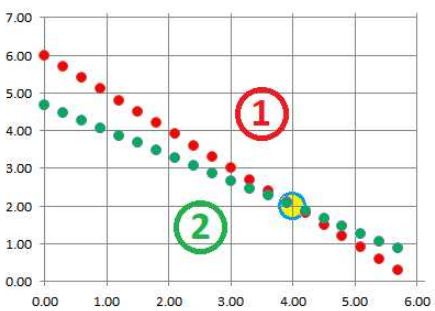
- x-axis $\cap$ y-axis
- $\{$ x-intercepts $\} \cap \{$ y-intercepts $\}$

Example 3.1.15: mixtures, revisited

Recall an example from Chapter 2: Is it possible to create, from the Kenyan coffee ($2 per pound) and the Colombian coffee ($3 per pound), six pounds of blend with a total price of $14? We produced two linear relations, each representing a line on the xy-plane:

$$\begin{cases} x + y = 6\,, \\ 2x + 3y = 14\,. \end{cases}$$

These are the lines:



We conclude that the *intersection* of the two lines is the point (x, y) that is the solution of the *system* of equations formed by these two equations. Just as before, when there are two equations, there are two sets:

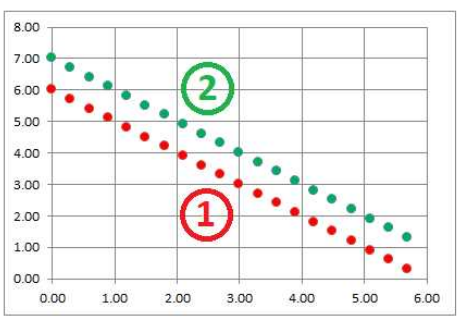
$$\begin{aligned} & \{(x,y):\ x+y=6 \quad \texttt{AND} \quad 2x+3y=14\} \\ = \ & \{(x,y):\ x+y=6\} \quad \cap \quad \{(x,y):\ 2x+3y=14\} \\ = \ & \{(4,2)\}\,. \end{aligned}$$

The intersection is a single point!

The existence of an intersection point tells us that it is *possible* to create such a blend! However, if the Colombian coffee is also priced at $2 per pound, there is no such mixture:

$$\{(x,y):\ x+y=6\} \cap \{(x,y):\ 2x+2y=14\} = \varnothing\,.$$

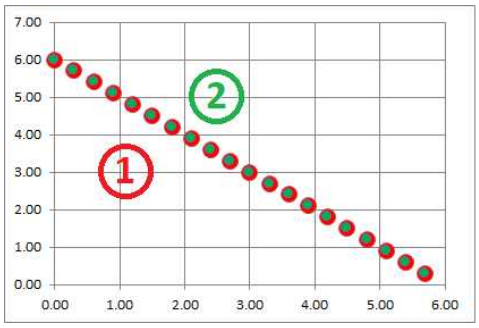
The lines are parallel:



The third possibility occurs when we change, in addition, the total price of the blend to $12:

$$\{(x,y):\ x+y=6\} \cap \{(x,y):\ 2x+2y=12\} = \{(x,y):\ x+y=6\}\,.$$

The lines merge:



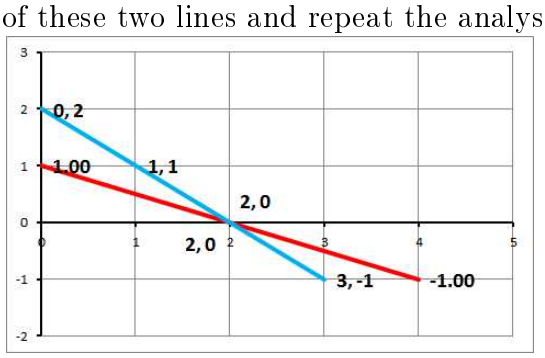
There are infinitely many possible mixtures.

Exercise 3.1.16

Carry out the computations and provide visualization for the second and third possibilities in the last example.

Exercise 3.1.17

Find algebraic representations of these two lines and repeat the analysis above:

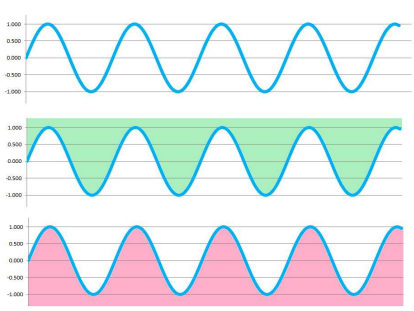


Exercise 3.1.18

Evaluate:

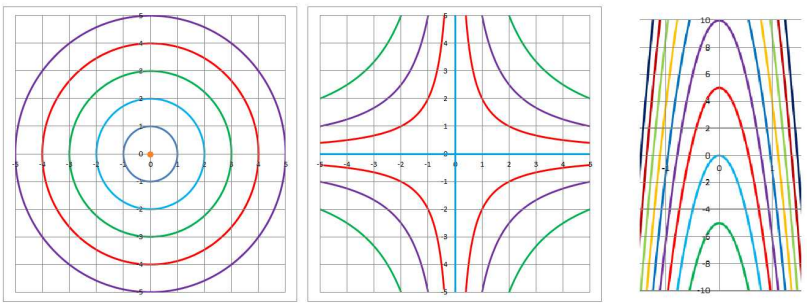
$$\{(x, y) : \ y \geq f(x) \ \} \cap \{(x, y) : \ y \leq f(x) \ \} .$$

Hint:



Example 3.1.19: graphs as subsets

With more room on the plane, the sets can overlap in a variety of ways. All *graphs* – of both relations and functions – are subsets of $\mathbf{R}^2$ (Chapter 2). For example, these are a few graphs of some relations, such as $x^2 + y^2 = k$, producing concentric circles (left):



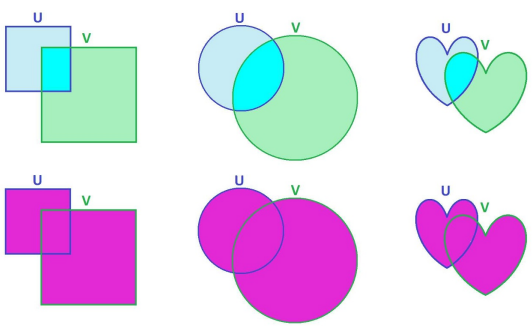
The pictures for the hyperbolas (middle) and parabolas (right) are similar. What do they have in common? We observe that:

1. The intersection of any two of them (for two different values of k) is empty.
2. The union is the whole plane.

Exercise 3.1.20

Prove the last statement.

More examples of these two operations for subsets of the plane are shown below:



These "generic" sets serve as illustrations of the idea of the union and the intersection.

3.2. Piecewise-defined functions

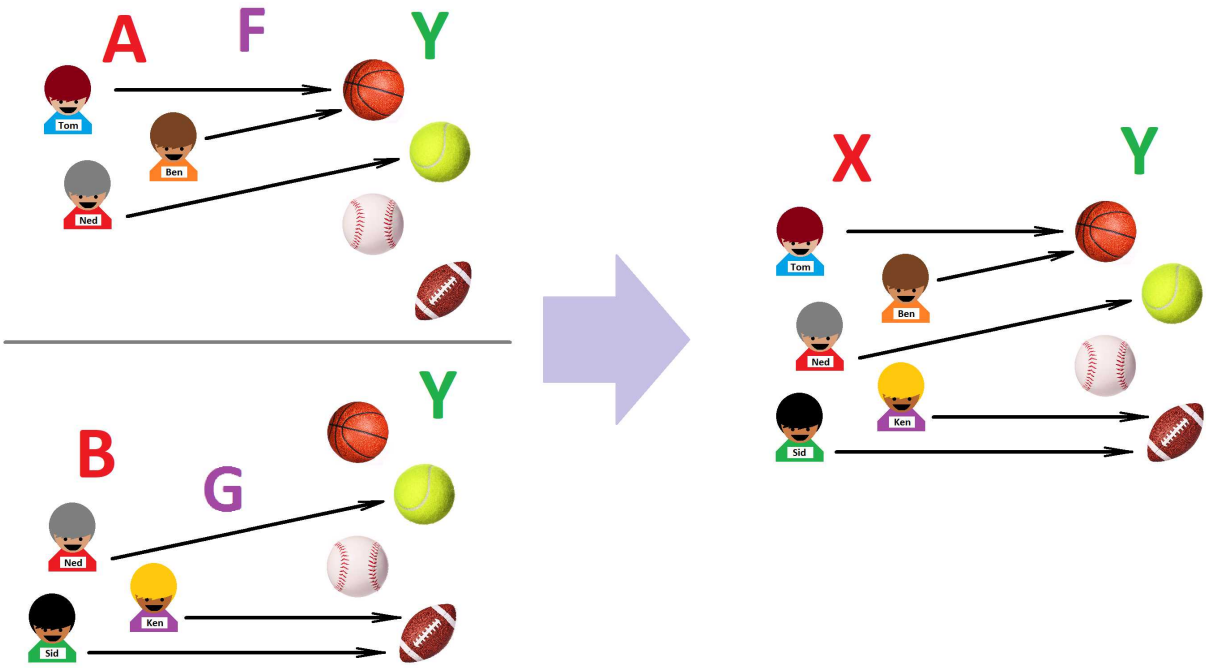
Back to our running example, suppose one person knows only the preferences of Tom, Ned, and Ben, i.e., some function

$$F : A = \{\text{Tom, Ned, Ben}\} \to Y\,;$$

and suppose another person knows only the preferences of Ben, Ken, and Sid, i.e., some other function

$$G : B = \{\text{Ben, Ken, Sid}\} \to Y\,.$$

Can the two pool their knowledge together in a meaningful way? This amounts to building a new function $H : A \cup B \to Y$ from the two old ones:

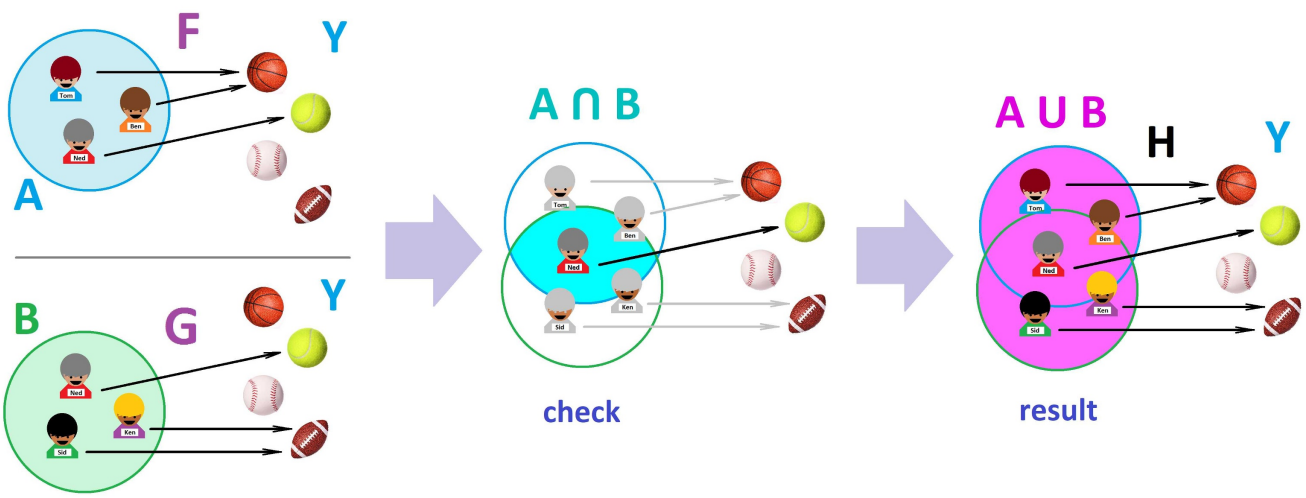


Of course, this is only possible if the data of former isn't in conflict with that of the latter (what does Ben prefer?). Otherwise we would end up with two arrows originating from the same input – not a function!

We have two observations:

1. We define a new function and its domain is the *union* of the domains of the two functions.
2. It is only possible as long as the values of the two functions on the *intersection* of their domains are the same.

We just merge the two together:



We confirm that the information about Ben matches (middle) and then we complete the merge (right).

The two domains become the two *pieces* of the domain of the new function.

Definition 3.2.1: piecewise-defined function

Let A and B be two subsets of a set X such that $X = A \cup B$ and let Y also be a set. Suppose also that $F : A \to Y$, $G : B \to Y$ are two functions. A function $H : X \to Y$ defined by the following rule:

- if x is in A, then $f(x) = F(x)$,
- if x is in B, then $f(x) = G(x)$,

is called the *piecewise-defined function* of F and G.

This is the flow chart of the new function:

$$H:\ x \ \to \ \begin{array}{l} \nearrow \text{if } x \text{ in } A \quad x \to \boxed{F} \to y \searrow \\ \searrow \text{if in } B \quad x \to \boxed{G} \to y \nearrow \end{array} \ y \ \to \ y$$

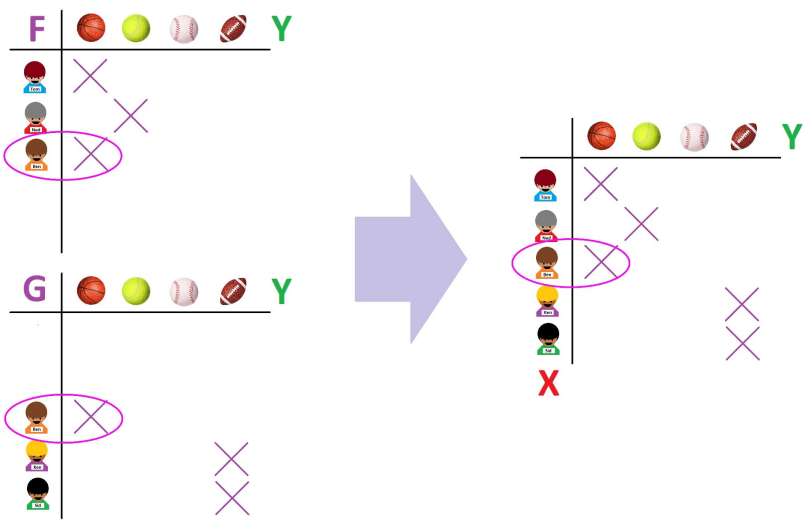
Since f is assumed to be a function, we are supposed to have one y for each x in the domain. Therefore, for f to exist, we must have the following:

$$F(x) = G(x) \quad \text{for every } x \text{ in } A \cap B\,.$$

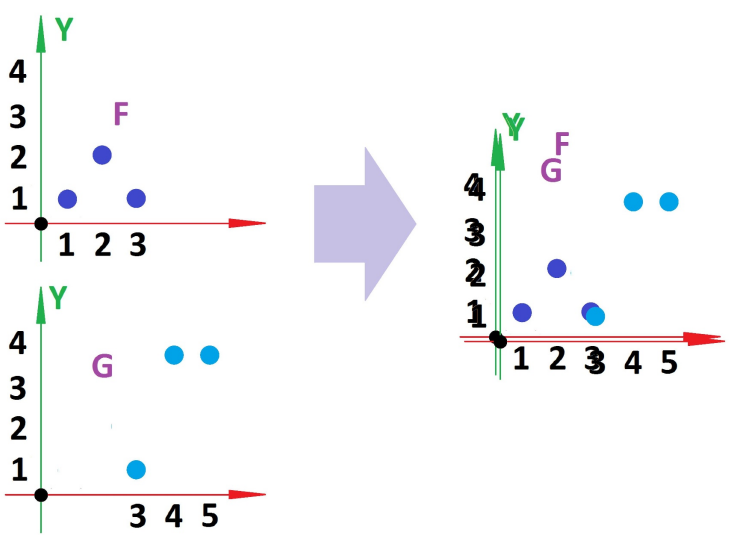
Exercise 3.2.2

What if the *codomains* are also different?

To build f from F and G, we can simply *merge the tables* of these functions:



We can also *superimpose the graphs* of these functions:



The algebraic notation for this function puts the formulas before the conditions:

Piecewise notation

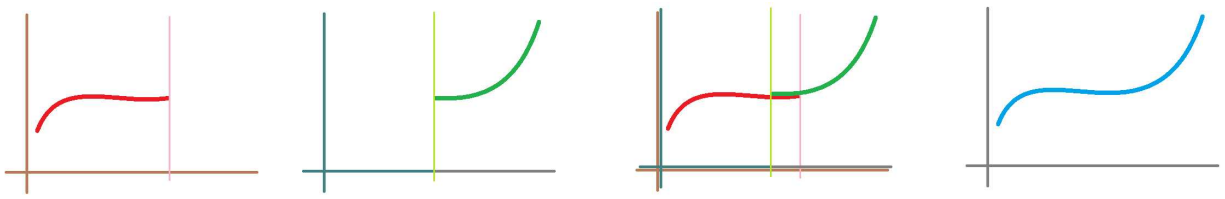
$$f(x) = \begin{cases} F(x) & \text{if } x \text{ in } A \\ G(x) & \text{if } x \text{ in } B \end{cases}$$

As it is often the case in algebra, the formula is to be read from right to left; i.e.,

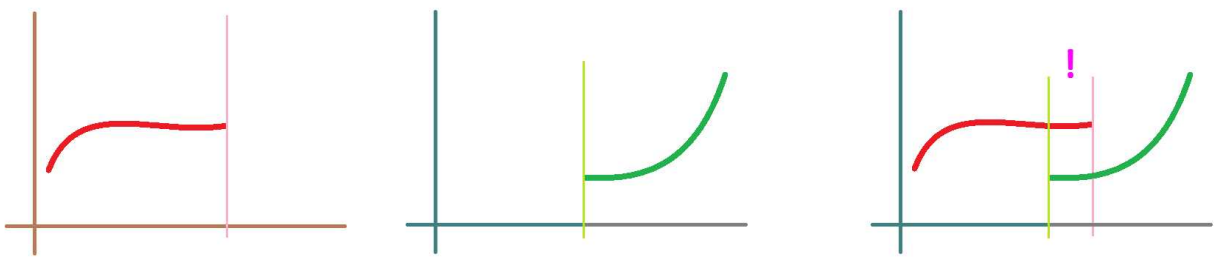
$$\begin{cases} \text{If } x \text{ in } A, & \text{then assign} & f(x) = F(x)\,. \\ \text{If } x \text{ in } B, & \text{then assign} & f(x) = G(x)\,. \end{cases}$$

Example 3.2.3: gluing graphs together

In the case of numerical functions, the graphs of two functions are "glued together" (or we may see "two wires soldered together") to create a new function, as follows:



The construction fails when the values don't match:



The *Vertical Line Test* is violated!

When there is no intersection of the domains,

$$A \cap B = \varnothing\,,$$

there can be no conflict! We can then easily glue together as many functions as we like. For any collection of sets $A_1, A_2, ...$, and functions $F_1, F_2, ...$ with $A_i \cap A_j = \varnothing$ for $i \neq j$, we use the following notation:

Piecewise-defined function

$$f(x) = \begin{cases} F_1(x) & \text{if } x \text{ in } A_1 \\ F_2(x) & \text{if } x \text{ in } A_2 \\ F_3(x) & \text{if } x \text{ in } A_3 \\ \dots & \dots \end{cases} \quad \text{means} \quad \begin{cases} \text{If } x \text{ in } A_1\,, & \text{then assign} & f(x) = F_1(x)\,. \\ \text{If } x \text{ in } A_2\,, & \text{then assign} & f(x) = F_2(x)\,. \\ \text{If } x \text{ in } A_3\,, & \text{then assign} & f(x) = F_3(x)\,. \\ \dots & \dots & \end{cases}$$

With multiple sets, the "pieces" of the domain are introduced as *conditions* to be satisfied.

Example 3.2.4: income tax

Suppose, hypothetically, that the tax code says:

1. If your income is less than \$10000, there is no tax to pay.
2. If your income is between \$10000 and \$20000, the tax rate is 10%.
3. If your income is over \$20000, the tax rate is 20%.

Let's build a function! Let x be the input (the income) and $y = f(x)$ output (the tax rate). We choose the domain to be $[0, +\infty)$ and we choose the codomain to be $[0, 1]$. The three "brackets" are the subsets of the domain. Since the values are different, we need to make sure that there is no overlap between the "pieces":

1. If $0 \leq x \leq 10000$, then $y = 0$.
2. If $10000 < x \leq 20000$, then $y = .10$.
3. If $20000 < x$, then $y = .20$.

Let's build a flowchart. Unlike most of the recent examples, there is a *fork on the road* this time. The

way we answer these questions about x determines the route for our computation:

$$f: \; x \;\to\; x \;\to\; \begin{array}{ll} \nearrow \text{ if } x\le 10000 & x \to \boxed{\text{choose } 0} \to y \searrow \\ \to \text{ if } 10000<x\le 20000 & x \to \boxed{\text{choose } .10} \to y \to y \to y \\ \searrow \text{ if } 20000<x & x \to \boxed{\text{choose } .20} \to y \nearrow \end{array}$$

The graph is on the left:



We also express the tax rate algebraically:

$$f(x) = \begin{cases} 0 & \text{if } x \le 10000\,, \\ .10 & \text{if } 10000 < x \le 20000\,, \\ .20 & \text{if } 20000 < x\,. \end{cases}$$

This formula, however, shouldn't be applied directly to calculate your tax bill; indeed, if your income rises from \$10,000 to \$10,001, your tax bill would rise from \$0 to \$1,000. This is the issue of "discontinuity" (discussed in Chapter 2DC-2). More realistically, the output of the formula represents a "marginal" tax rate: the rate you apply to that part of the income that lies within the bracket. These three numbers are then added together to produce your tax bill.

Exercise 3.2.5

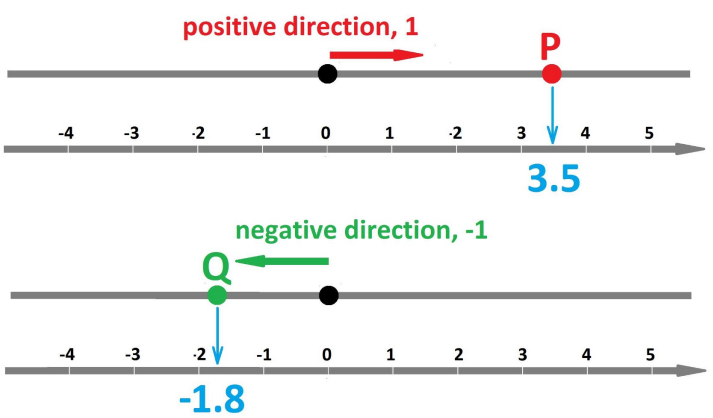
Find the formula of the tax bill. Hint: Its graph is shown above on the right. What are the slopes?

Some very important functions are presented in the piece wise manner.

The following function is meant to give the *direction* from the origin to the location given by the input number:

- 3 is to the right of 0, so the direction is positive; and we say "it's 1";
- -5 is to the left of 0, so the direction is negative; and we say "it's -1".

Another example is shown below:



Definition 3.2.6: sign function

The *sign function* (or the signum) is computed according to the following rule:

- if $x < 0$, then assign $y = -1$,
- if $x = 0$, then assign $y = 0$,
- if $x > 0$, then assign $y = 1$.

It is denoted by

$$y = \operatorname{sign}(x)$$

We think of a function as a sequence of steps, but this is an example of one with a fork:

$$y = \operatorname{sign}(x): \quad x \quad \to \quad x \quad \begin{array}{l} \nearrow \text{ if } x<0 \\ \to \text{ if } x=0 \\ \searrow \text{ if } x>0 \end{array} \quad \begin{array}{l} \to \boxed{\text{choose } -1} \\ \to \boxed{\text{choose } 0} \\ \to \boxed{\text{choose } 1} \end{array} \quad \begin{array}{l} \to y \searrow \\ \to y \to y \to y \\ \to y \nearrow \end{array}$$

Note how x's have disappeared! Below is an algebraic expression of the flowchart above:

$$\operatorname{sign}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

In other words, the function strips the input number of its magnitude and what's left is its sign, which is, in a way, its direction.

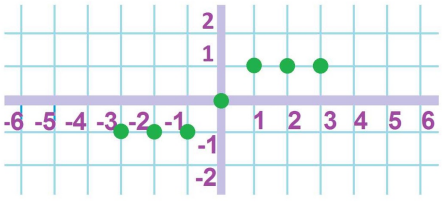
In this example, note how there are three doors with specific values behind each and the only job of x is to point us to the right door.

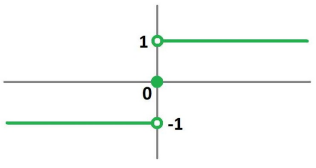
It is clear that the implied domain of this function is $X = \mathbf{R} = (-\infty, +\infty)$.

Since the only possible values are listed in the description of the function, the image is the set of these three, $V = \{-1, 0, 1\}$. We see that also in the table of values of the sign function:

x	-3	-2	-1	0	1	2	3
y	-1	-1	-1	0	1	1	1

This is the graph built from this data (left):





This is not enough, however, if we want to see what the graph is like around 0; we insert more points into the table (below) and then plot (above right):

x	$-.8$	$-.6$	$-.4$	$-.2$	0	$.2$	$.4$	$.6$	$.8$
y	-1	-1	-1	-1	0	1	1	1	1

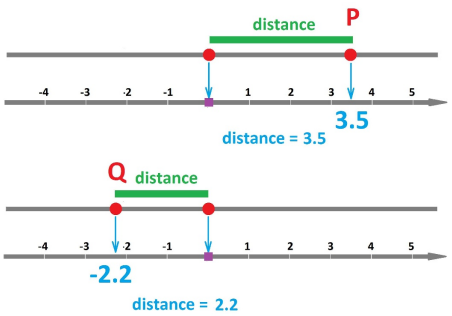
The outputs continue to be repeated! This is an indication that the graph is horizontal in those areas.

Exercise 3.2.7

Does the graph have a symmetry? What are the slopes?

As we plot those points, we realize that we can't connect them into a "single curve". If the function were to represent the location as a function of time, we would teleport. (This issue is discussed in Chapter 2DC-2.)

The next function is meant to give the *distance* from the location given by the input number to the origin:



In particular, both 3 and -3 are 3 units from the origin. The result is always positive (except for 0 itself). We commonly say that "the minus sign is dropped".

Warning!

We can't literally "drop the sign" of $-x$ because x itself might be negative.

How else do we make a negative number positive? We multiply it by -1!

The following will re-appear many times.

Definition 3.2.8: absolute value function

The *absolute value function* icomputed according to the following rule:

- if $x < 0$, then assign $y = -x$,
- if $x = 0$, then assign $y = 0$,
- if $x > 0$, then assign $y = x$.

It is denoted by

$$y = |x|$$

This is the formula:

$$|x| = \begin{cases} -x & \text{if } x < 0\,, \\ 0 & \text{if } x = 0\,, \\ x & \text{if } x > 0\,. \end{cases}$$

And this is its flowchart representation:

$$y = |x| : \quad x \;\to\; \begin{array}{x} \nearrow \text{if } x<0 \\ \to \text{if } x=0 \\ \searrow \text{if } x>0 \end{array} \begin{array}{l} x \to \boxed{\text{multiply by } -1} \to y \searrow \\ x \to \boxed{\text{choose 0}} \to y \\ x \to \boxed{\text{pass it}} \to y \nearrow \end{array} \to y \to y$$

Exercise 3.2.9

Explain the alternative method of computing this function:

- If $x \leq 0$, then $y = -x$.
- If $x \geq 0$, then $y = x$.

Exercise 3.2.10

Present a flowchart for this definition.

In other words, the function does the opposite of what the sign function does; it strips the input number of its sign, or direction, and what's left is its magnitude.

Exercise 3.2.11

Show that the two functions above are related by the following identity:

$$x = \text{sign}(x) \cdot |x| \,.$$

It is clear that the domain of this function is $X = \mathbf{R} = (-\infty, +\infty)$.

To find the image (i.e., the range of values), we take into account these two facts:

- The only possible values are non-negative (by design).
- Every non-negative number is equal to its own absolute value.

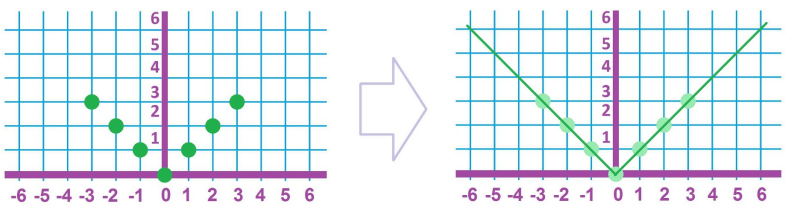
Therefore, the image is the set of all non-negative numbers, $[0, +\infty)$. We see that also in the table of values of the absolute value function, the y-values:

x	-3	-2	-1	0	1	2	3
y	3	2	1	0	1	2	3

Exercise 3.2.12

Does the graph have a symmetry? What are the slopes?

As we plot more and more of those points (left), we realize that unlike $y = \text{sign}(x)$, we *can* connect them into a "single curve" (right):



However, this V-shaped curve is made of parts of two straight lines, $y = x$ and $y = -x$, without a smooth transition. If the function were to represent the location as a function of time, the change of direction would be abrupt or even instantaneous. (This issue is discussed in Chapter 2DC-3.)

Exercise 3.2.13

What symmetry do you see in the graph?

Another important piecewise-defined function is the following:

Definition 3.2.14: integer value function

The *integer value function* is defined as the largest integer y less than or equal to x; it is denoted by

$$y = [x]$$

For example, we have:

$$[5] = 5, \;\; [3.1] = [3.99] = 3, \;\; [-4.1] = -5 \,.$$

Given an input of x, let's describe a procedure for computing the output of this function. We have for a given x:

- Step 1: Determine whether x is an integer or not.
- Step 2: If x is an integer, then set $y = x$; otherwise, decrease from x until meet an integer, and set y equal to that integer.

Exercise 3.2.15

Present a flowchart for this definition and a formula.

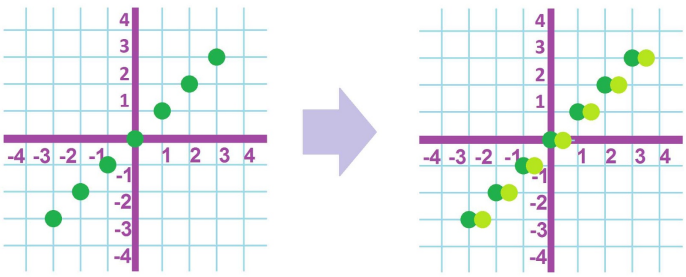
It is clear that the domain of this function is all reals $\mathbf{R} = (-\infty, +\infty)$.

To find the image, we take into account these two facts:

- The only possible values are integers (by design).
- The integer value of every integer is that number.

Therefore, the image is the integers $\mathbf{Z} = \{..., -3, -2, -1, 0, 1, 2, 3, ...\}$.

This is not enough, however, in order to see what the graph is. All we see is just a sequence of points on the line $y = x$:



We need to insert more points (in the middle of the intervals):

x	-3	-2.5	-2	-1.5	-1	$-.5$	0	$.5$	1	1.5	2	2.5	3
y	-3	-3	-2	-2	-1	-1	0	0	1	1	2	2	3

The output is the same to the one of the adjacent integer! This is an indication that the graph is horizontal in those areas.

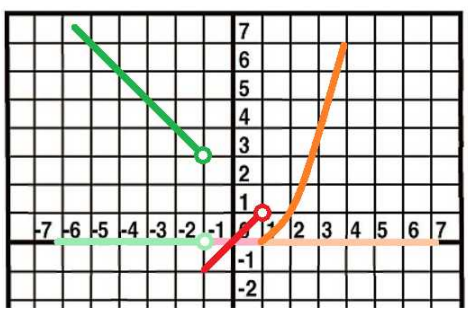
Exercise 3.2.16

Does the graph have a symmetry? What are the slopes?

As we plot these points, we realize that we can't connect them into a "single curve", just as the sign function.

Example 3.2.17

Note that the "pieces" of a piecewise-defined function refer to the subsets of the *domain* and they might cover the whole real line without intersections. Meanwhile, the corresponding pieces of the *graph* don't have to fit together:



Exercise 3.2.18

Write a formula for the above graph.

Example 3.2.19: plotting piecewise

Let's plot the graph of the following piecewise-defined function:

$$f(x) = \begin{cases} 3 - \frac{1}{2}x & \text{for } x \leq 2\,, \\ 2x - 5 & \text{for } x > 2\,. \end{cases}$$

Since there is no intersection of the two sets, there can be no conflict.

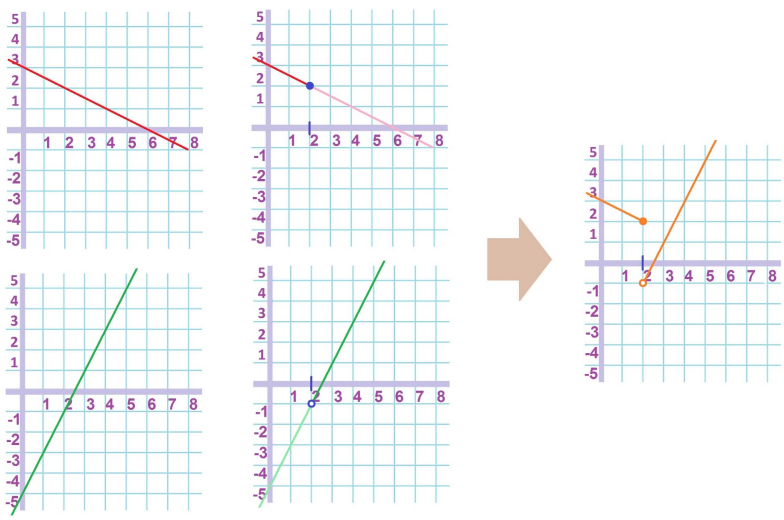
The idea is to consider the formula as two unrelated functions:

	formula	domain
1.	$y = 3 - \frac{1}{2}x$	$x \leq 2$
2.	$y = 2x - 5$	$x > 2$

We now plot them separately. We follow these steps:

1. Plot $y = 3 - \frac{1}{2}x$,.
2. Plot $y = 3 - \frac{1}{2}x$ with domain $x \leq 2$ (new function).
3. Plot $y = 2x - 5$.
4. Plot $y = 2x - 5$ with domain $x > 2$ (new function).
5. Merge the plots from (2) and (4).

These are the results:



Exercise 3.2.20

Make a flowchart, write a formula, and then plot the graph of the function $y = f(x)$, where x is the time in hours and $y = f(x)$ is the parking fee over x hours, which is computed as follows: free for the first hour, \$1 per every full hour for up to 4 hours, and a flat fee of \$5 for anything longer.

Exercise 3.2.21

Make a hand-drawn sketch of the graph of the function

$$f(x) = \begin{cases} -3 & \text{if } x < 0\,, \\ x^2 & \text{if } 0 \leq x < 1\,, \\ x & \text{if } x > 1\,. \end{cases}$$

Example 3.2.22: domain vs. flowchart

The flowcharts also help to visualize what happens inside the black box of a function when input is outside the domain:

$$y = 1/x: \quad x \to \quad x \quad \begin{array}{l} \nearrow \text{ if } x \neq 0 \quad x \to \boxed{\text{take its reciprocal}} \to y \searrow \\ \searrow \text{ if } x = 0 \quad x \to \boxed{\text{STOP!}} \end{array} \quad y \to y$$

Exercise 3.2.23

Make a flowchart as above for $f(x) = \sqrt{x}$.

3.3. Numerical functions are transformations of the line

...and vice versa.

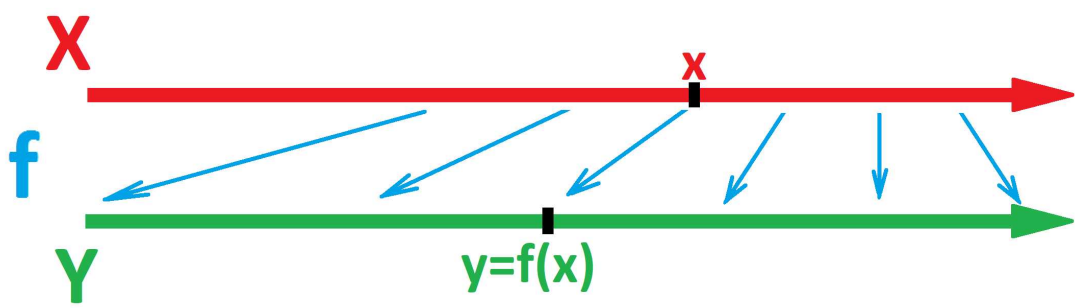
When we face a numerical function $y = f(x)$ given to us without any prior background, it is a good idea to create a *tangible representation* for it. The two main ways we are familiar with are these:

- We think of the function as if it represents *motion*: x time, y location.
- We plot the *graph* of the function on a piece of paper.

Let's set these aside for now and consider a third approach:

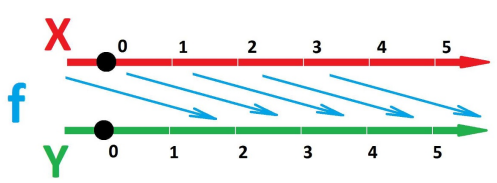
- We think of the function as a *transformation* of the real line.

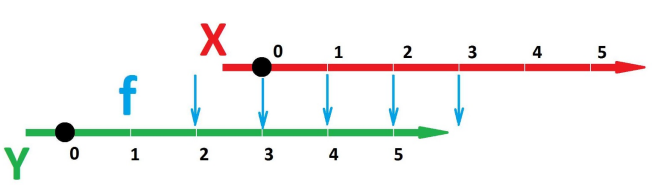
First, let's go back to representing a numerical function as correspondence between the x-axis and the y-axis:



As you can see, these are the same axes we use to plot the graph, but they are arranged *parallel* to each other instead of perpendicular. The arrows indicate where each x lands on the y-axis. In fact, they suggest what happens to the *whole* x-axis!

Let's consider something specific:



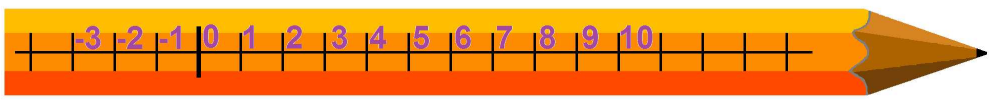


We ask: What has happened to the x-axis under this transformation?

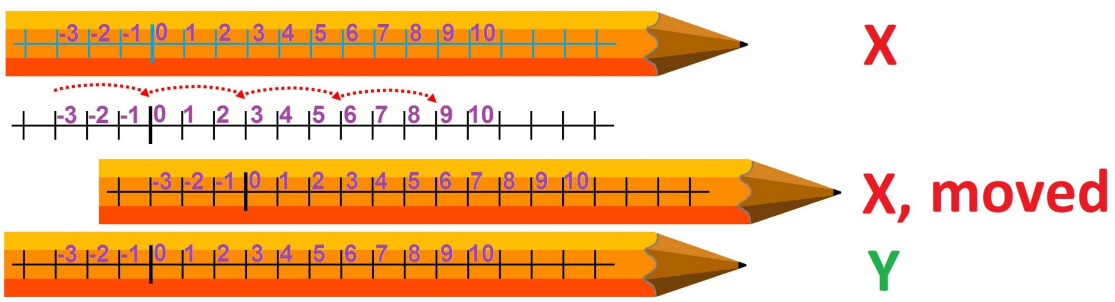
Warning!

Above, you see two ways to interpret the function: (1) arrows are between the x-axis and the intact y-axis, or (2) we move the y-axis so that $y = f(x)$ is aligned with x. The approaches give two different, even opposite, answers to the question. We will follow the former.

To make this more tangible, we will think as if the whole x-axis is drawn on a pencil:



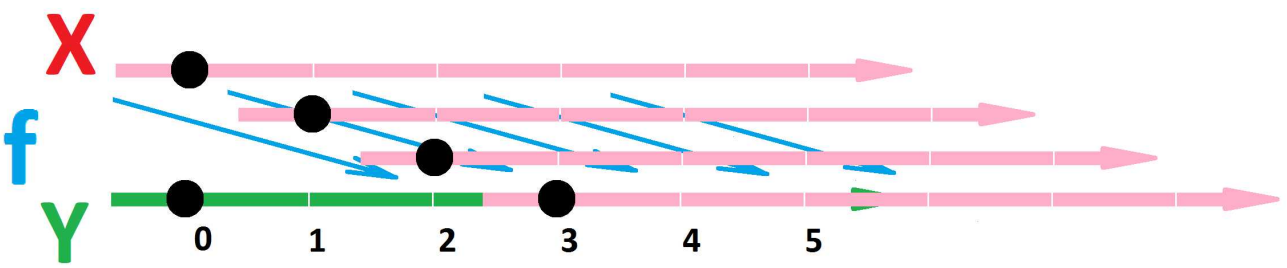
The transformation we are starting with is a *shift*. We simply slide the pencil horizontally. Furthermore, there is another pencil to be used for reference. The markings (i.e., its coordinate system) on the second pencil show the new locations of the points on the first. As an example, a shift of 3 units to the right is shown below:



Generally:

- When shifted $s > 0$ units right, point x becomes $x + s = y$.

This is where the algebra comes from:



Warning!

Replace "right" with "up" if the axes are aligned vertically; it's the positive direction that matters.

In the meantime, what about shifting left? We have the following:

- When shifted $s > 0$ units left, point x becomes $x - s$.

Of course, we can combine the two statements:

- When shifted $s > 0$ units right/left, point x becomes $x \pm s$.

Instead, we should simply allow s to be *negative*. Then we can interpret the former statement to include

the latter if we understand "s units left" as "$-s$ units right". This is better:

$$x \xrightarrow{\text{right by } s} x+s$$

In other words, we have the following function:

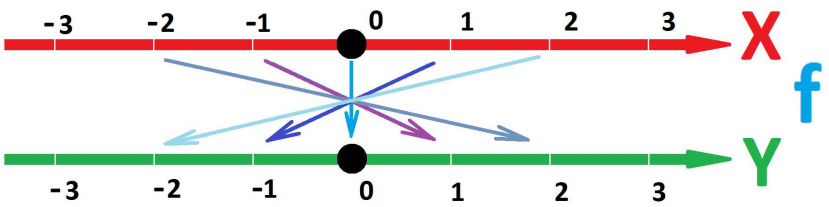
$$y = f(x) = x + s\,.$$

Of course, $s = 0$ means that there is no change.

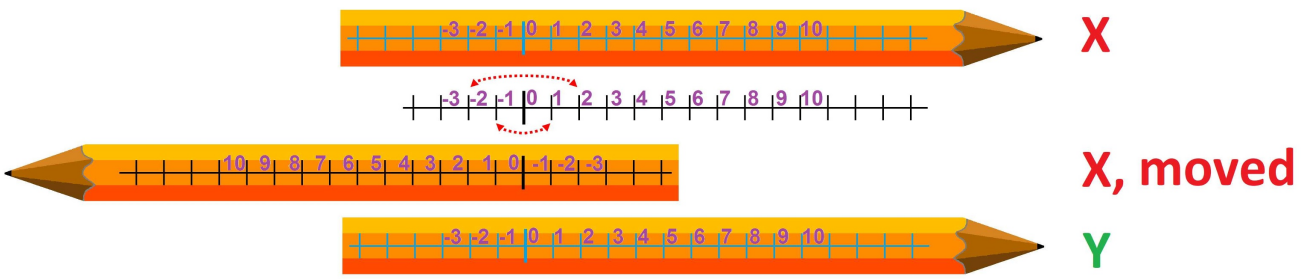
Next, let's consider a transformation produced by the function below:

$$x \xrightarrow{\text{flip}} -x$$

We plot the arrows to show where each point goes:



To understand what is happening to the whole x-axis, we try to imagine what happens to our pencil. We lift, then *flip* the pencil with x-axis on it, and place it next to another such pencil used for reference:

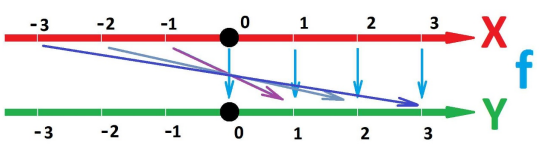


The transformations that don't distort the line are called *rigid motions*. Examples of motions are the left shift and the right shift (the magnitude may vary) and the flip (the center may vary).

Example 3.3.1: fold

Going in the opposite direction, we might want to find a formula for a transformation we describe verbally. A different kind of transformation is a *fold*. In order to visualize it, we can't put the axis on a pencil anymore! Let's try a piece of wire:



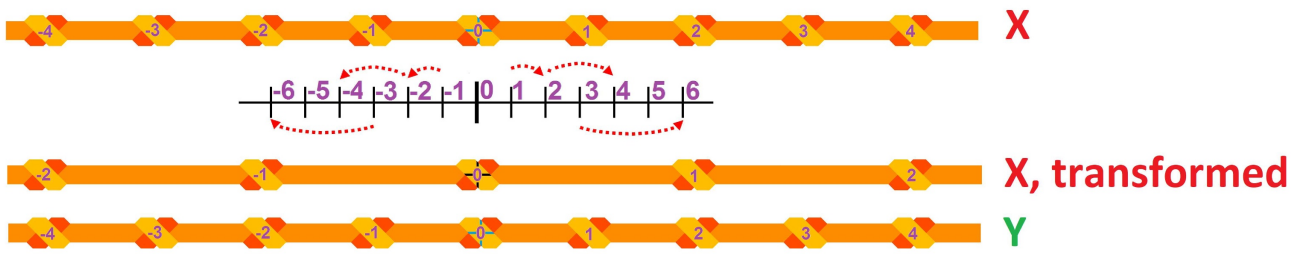


The arrow diagram on the right depicts how a half of the x-axis flips and the other stays put. What is this function? It is easy to see the range, $[0, +\infty)$, because that's where output land. This is the absolute value function, $y = |x|$!

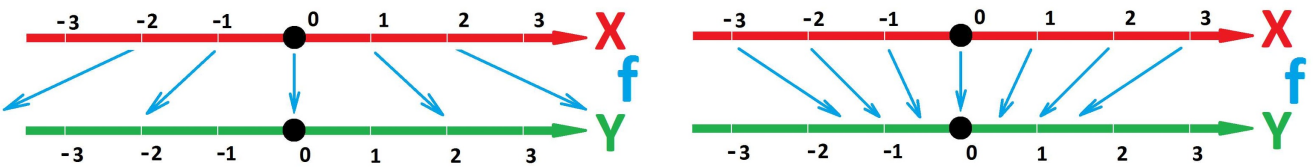
Another basic transformation is a *stretch*. The line isn't on a pencil or a wire anymore! It is a rubber string (with knots to mark locations):



We imagine that we grab it by the ends (at infinity?) and pull them apart in such a way that the center O doesn't move. For example, a stretch by a factor of 2 is shown below:



This is indeed a uniform stretch because the distance between *any* two points doubles (left):



The transformation is simply $y = 2x$.

Exercise 3.3.2

What about the second transformation shown?

Exercise 3.3.3

What does $f(x) = -2x$ do to the x-axis?

Warning!

If you stretch something, it becomes bigger, and if you shrink something, it becomes smaller – unless it's infinite!

Generally:

- When the line is stretched by a factor $k > 1$, point x becomes $x \cdot k$.

In the meantime, what about a *shrink*? We have the following:

- When the line is shrunk by a factor $k > 1$, point x becomes x/k.

Of course, in order to combine the two statements, we should allow k to be *less than* 1. Then we can interpret the former statement to include the latter if we understand "stretched by a factor k" as "shrunk by a factor $1/k$". This is how we can describe it:

$$x \xrightarrow{\text{stretch by } k} x \cdot k$$

Of course, $k = 1$ means that there is no change. In other words, this is the function:

$$y = f(x) = x \cdot k\,.$$

Exercise 3.3.4

What is the meaning of $|m|$ in the transformation given by $f(x) = mx + b$?

Exercise 3.3.5

What is the meaning of b in the transformation given by $f(x) = mx + b$?

Example 3.3.6: coloring numbers

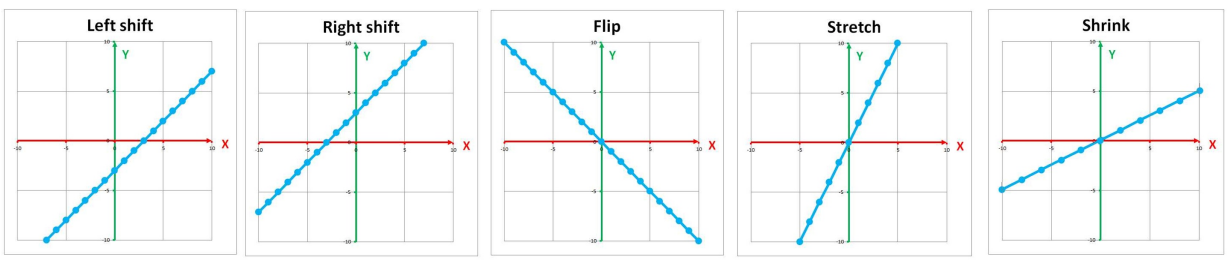
We will often color numbers according to their values. This is the summary of the functions we have considered:



One can compare the colors of X to those of X as it lands on y and reach the following conclusions:

- The colors are clearly shifted in the first two images.
- In the third, the blue and the red are interchanged.
- In the fourth, the colors don't change on Y as fast as on X, while in the fifth, they are changing faster.

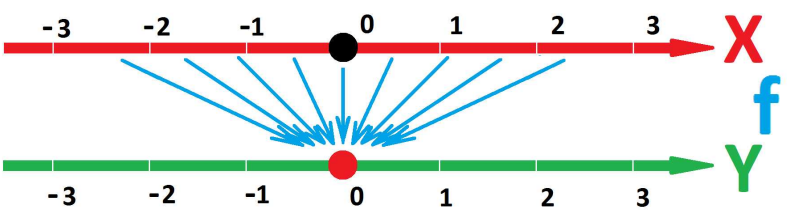
We also plot the graphs of these functions below:



We can, therefore, classify the transformations based on the slopes of their graphs:

- shift (rigid motion), slope 1;
- flip (rigid motion), slope -1;
- stretch by 2, slope 2;
- shrink by 2, slope 1/2.

A very simple, but important, class of function is the *constant functions*, $f(x) = c$. What does this function do to the x-axis? There is only one output:



The real line is shrunk to a single point; we can call this transformation *collapse*.

$$x \xrightarrow{\text{collapse}} c$$

Of course, one may see it as an extreme case of a stretch/shrink, with a factor $k = 0$. The slope of the graph is, of course, 0.

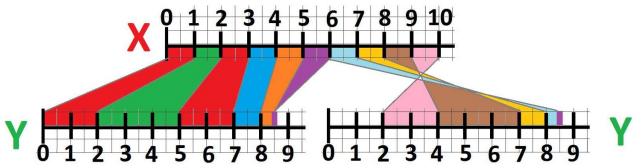
In general, it is typical to have different stretching factors at different locations.

Example 3.3.7: transformation from list of values

Let's consider this function given by its values:

x	0	1	2	3	4	5	6	7	8	9	10
y	0	2	5	7	8	8.5	8.5	8	7	4	2

We assume that the function continues between these values in a linear fashion. For example, the interval $[0, 1]$ is mapped to $[0, 2]$ linearly ($y = 2x$), the interval $[1, 2]$ to $[2, 5]$, etc. The 1-unit segments on the x-axis are stretched and shrunk at different rates, and the ones beyond $x = 6$ are also flipped over (left):





We also plot the graph of this function on the right; the stretching factors become the slopes! So, at its simplest, a non-linear function is a combination of linear patches.

Exercise 3.3.8

Compute the difference quotient for this function, i.e., the slopes of these patches. What is the relation of each slope to what happens to the segment?

Exercise 3.3.9

Represent the following function given by its values as a transformation:

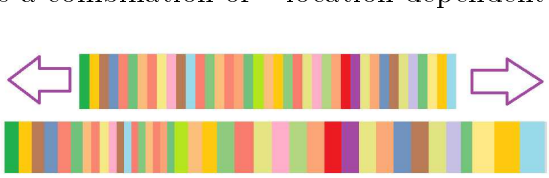
x	0	1	2	3	4	5	6	7	8	9	10
y	6	5	4	3	3	3	4	5	6	6	6

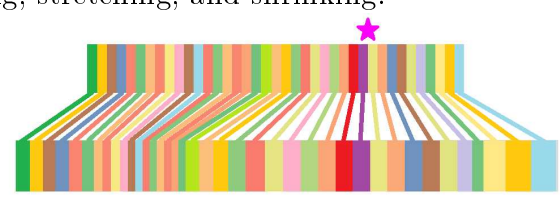
Make up your own functions and then represent them as transformations. Repeat.

Furthermore, it is also possible, and likely, for a stretch factor (or rate) of a function to vary *continuously* throughout the domain of a function. (This issue is addressed in Chapter 2DC-3.)

Example 3.3.10: fixed point

This is a combination of – location dependent – shifting, stretching, and shrinking:





Among all this complexity, the point marked with star is fixed.

Exercise 3.3.11

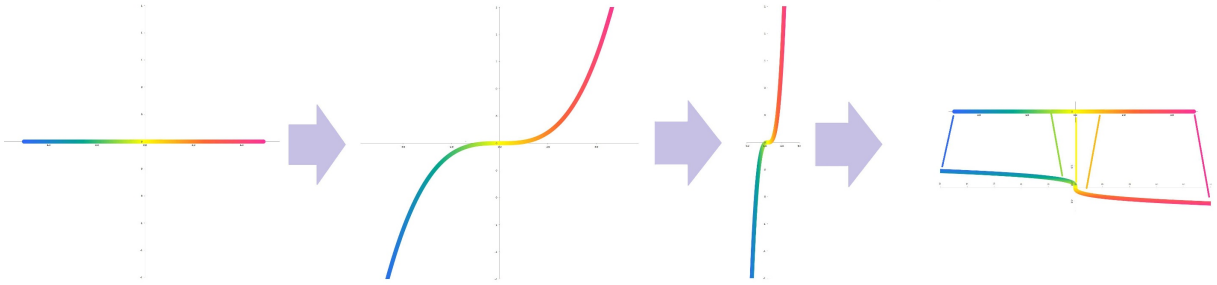
Sketch the graph of the function that represents the above transformation.

In summary, an abstract function $y = f(x)$ has been given three *tangible representations* and now we also have three tangible interpretations of its rate of change:

	Function is seen as	Its rate of change is
1.	location	velocity
2.	graph	slope of the tangent line
3.	transformation	stretch/shrink rate

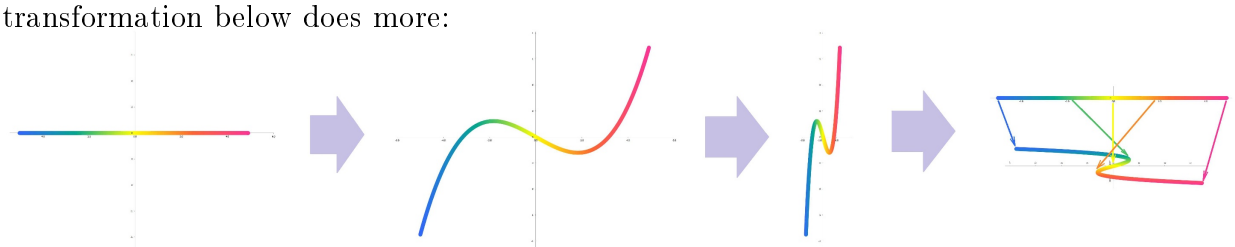
Example 3.3.12: graphs of transformations

Still, the most common way to visualize a function will remains plotting its *graph*. The picture below shows how the two approaches come together. The transformation of the domain into the codomain performed by the function can be seen in what happens to its graph:



First, we take the x-axis as if it is a (colored) rope and lift it vertically to the graph of the function, and second, we push it horizontally to the y-axis (practically, we just shrink the image). At the end, we can see what has happened to the whole line by pointing out what has happened to these three intervals: stretching at the ends and shrinking in the middle.

The transformation below does more:



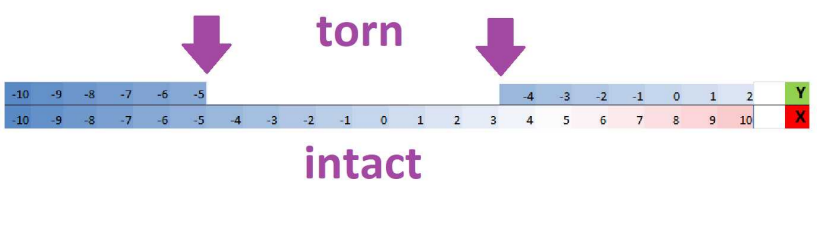
There is also a flip in the middle with folds.

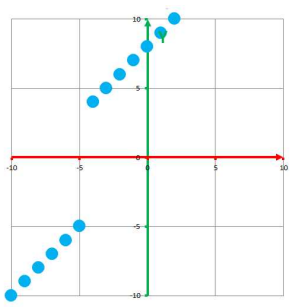
Exercise 3.3.13

Draw your own graph and then interpret the function as a transformation. Repeat.

Example 3.3.14: tearing and discontinuity

In general, a transformation can *tear* this rope. We can see how two numbers, 4 and 5, that are close to each other on the x-axis, are now far apart on the y-axis (right):

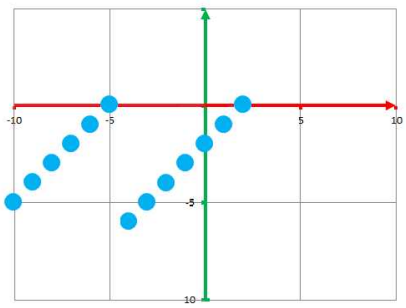




We can also see a gap in the graph (right). (The issue of "continuity" is addressed in Chapter 2DC-2.)

Exercise 3.3.15

Describe the function the graph of which is given below as a transformation:

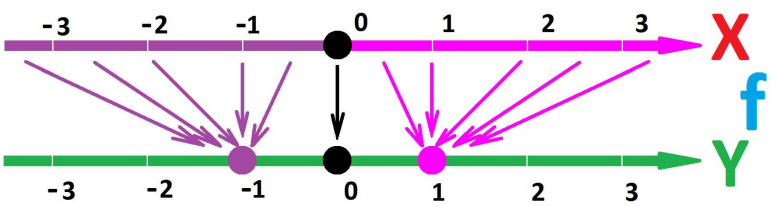


Exercise 3.3.16

Provide the graph of a function (a transformation) that tears this rope and then: (a) pulls the two halves apart, or (b) overlaps their ends.

Example 3.3.17: sign function

The sign function collapses the x-axis to three different points on the y-axis:

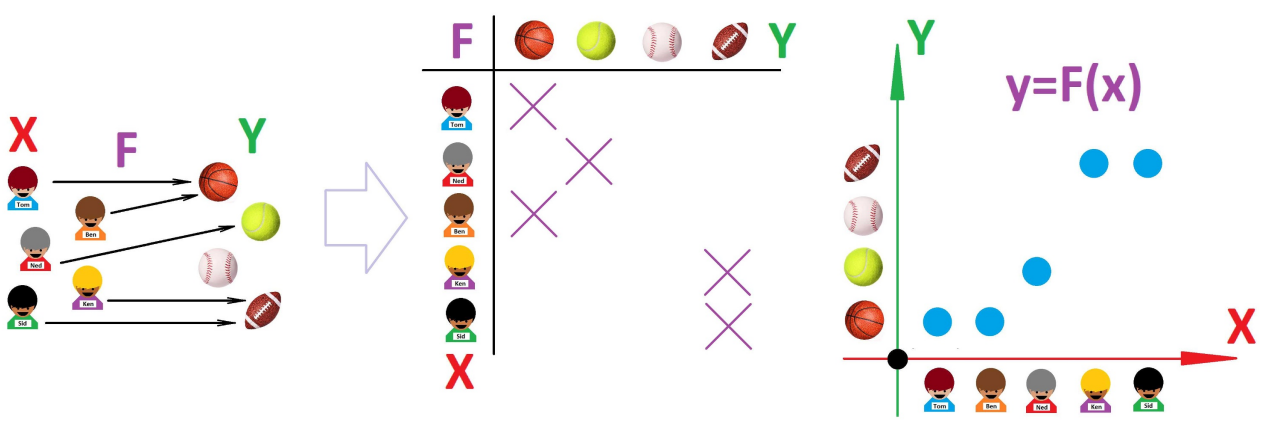


Exercise 3.3.18

What does the integer value function $y = [x]$ do to the x-axis?

3.4. Functions with regularities: one-to-one and onto

We go back to our example of a function that assigns to each boy a ball to play with. In order for this to be a function from the former to the latter, the table of this relation must have exactly one mark in each *row*:

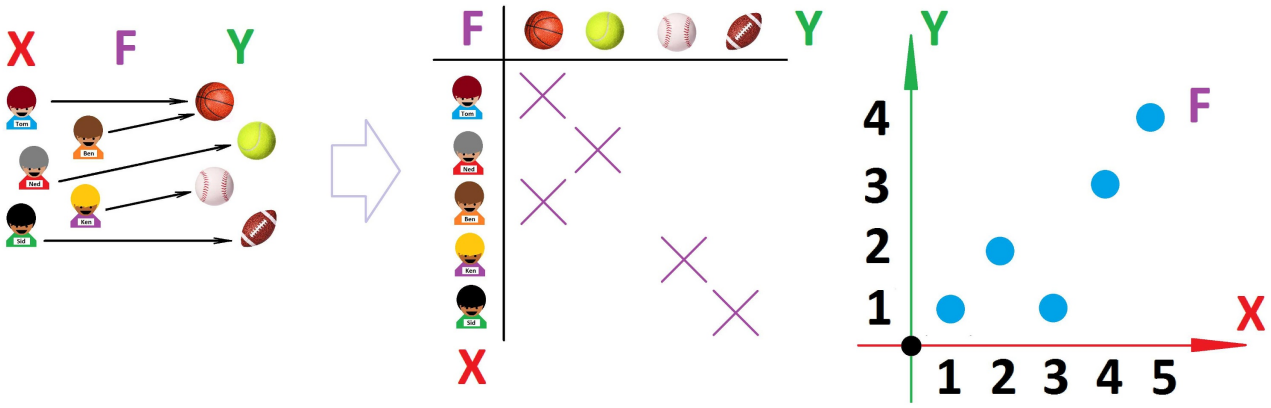


It does. But what about the *columns*?

If we further study this function, we might notice two different but related "irregularities".

First, no one seems to like baseball! There is no arrow ending at the baseball, and its column has no marks.

Let's modify the function slightly: Ken changes his preference from football to baseball. Then, there is an arrow for each ball, and all columns in the table have marks:



In the graph, there is a dot corresponding to each horizontal line.

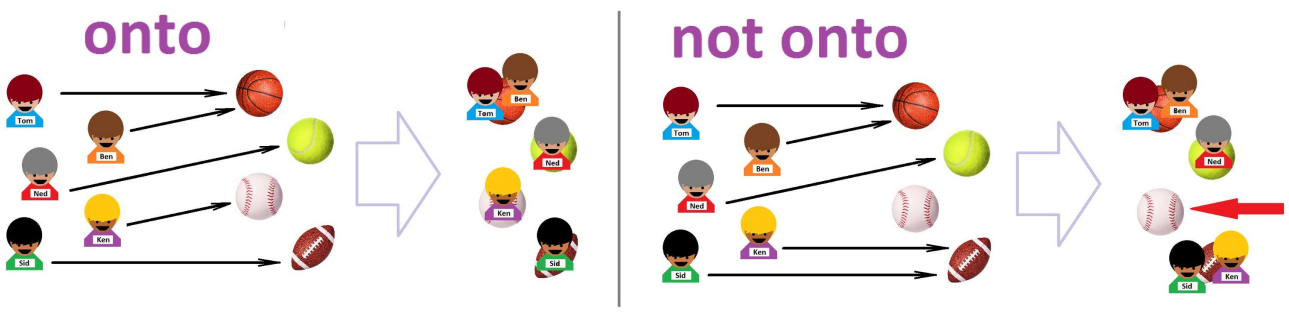
The following concept is crucial.

Definition 3.4.1: onto function

A function $F : X \to Y$ is called *onto* when there is an x for each y with $F(x) = y$.

In other words, no potential output is "wasted".

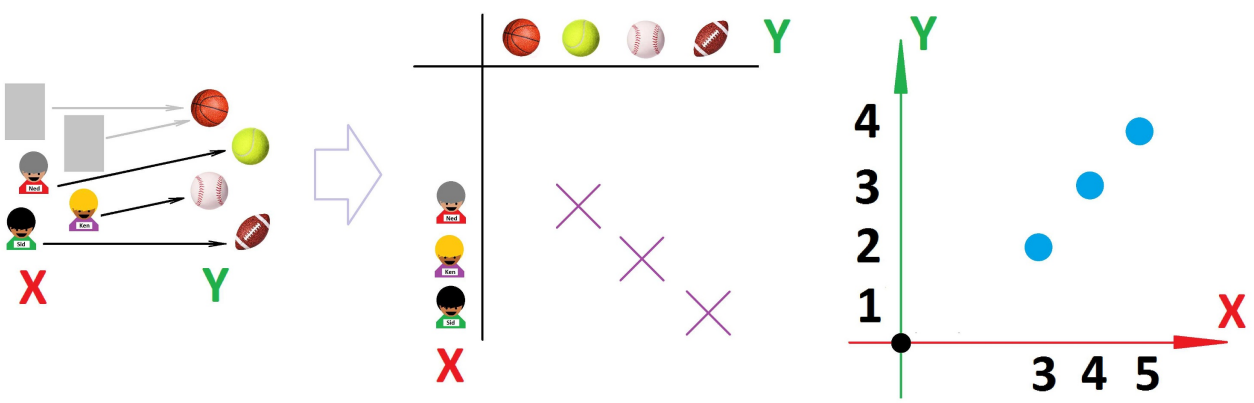
Below, the reason for this terminology is explained:



We start with an element of X and bring it – along the arrow – to the corresponding element of Y. The function is onto if all elements of Y are *covered*.

Second, both Tom and Ben prefer basketball! The two arrows converge on the basketball, and we can also see that its column has two marks. We note the same about the football.

The above function is modified: Tom and Ben have left. Then, no two arrows converge on one ball, and no column has more than one mark:



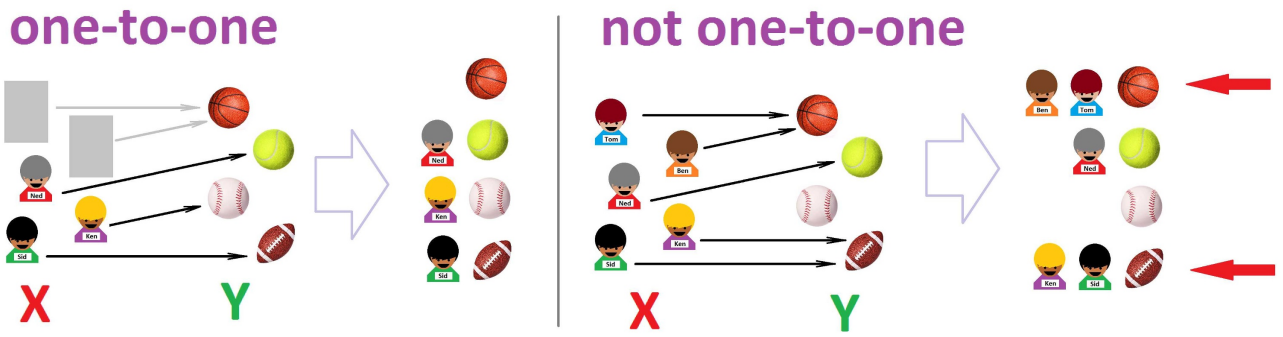
In the graph, there can only be one dot, or none, corresponding to each horizontal line.

The following concept is crucial.

Definition 3.4.2: one-to-one function

A function $F : X \to Y$ is called *one-to-one* when there is at most one x for each y with $F(x) = y$.

Below, the reason for this terminology is explained:



We start with an element of X and bring it – along the arrow – to the corresponding element of Y. The function is one-to-one if every element of Y is covered *only once*, if at all.

In summary, the two concepts are not about how many arrows *originate* from each x – it's always one – but about how many arrows *arrive* at each y. The logic of the two definitions is quite different:

- Onto: `FOR EACH` x `THERE IS` y such that $y = F(x)$.
- One-to-one: `IF` $F(x_1) = F(x_2)$ `THEN` $x_1 = x_2$.

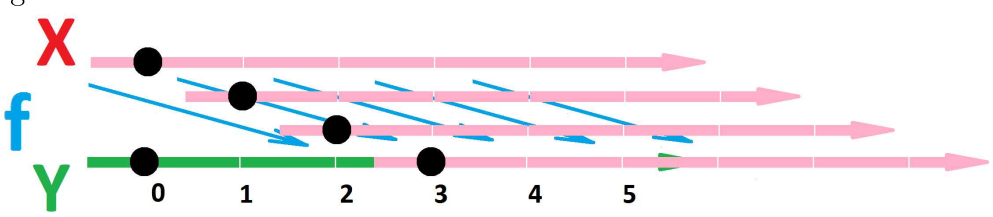
Now *numerical functions*. These functions are represented by their transformations and by their graphs. We will follow these two approaches to discover whether a function satisfies one of the two definitions. The following will be our assumption:

▶ The codomain is the set of all reals $\mathbf{R}$.

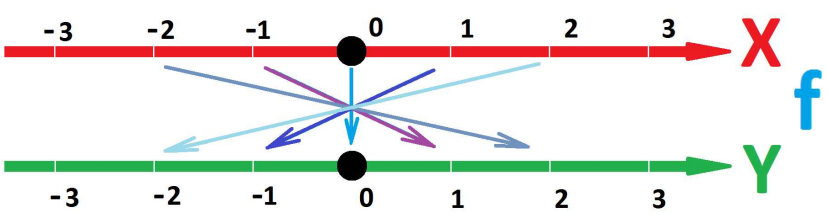
Example 3.4.3: basic transformations

Let's consider the transformations from the last section. Their descriptions – and illustrations – tell the whole story. We just need to look at how the arrows arrive at the y-axis.

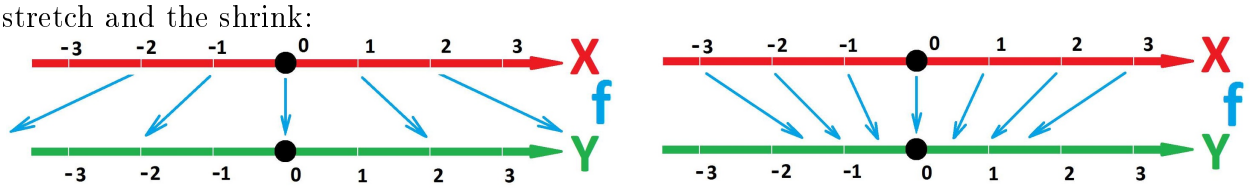
The left and right shifts:



The flip:

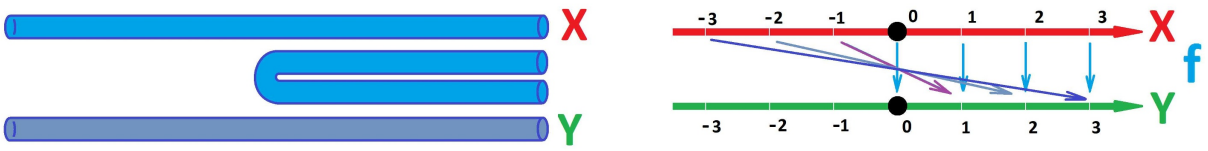


The stretch and the shrink:

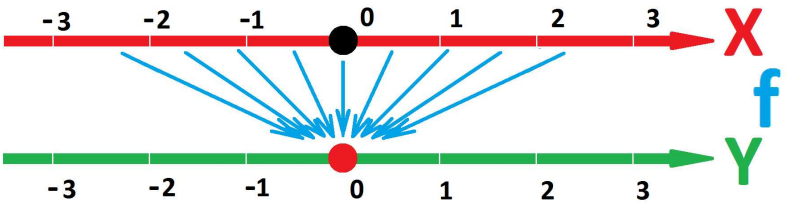


They are all both one-to-one and onto!

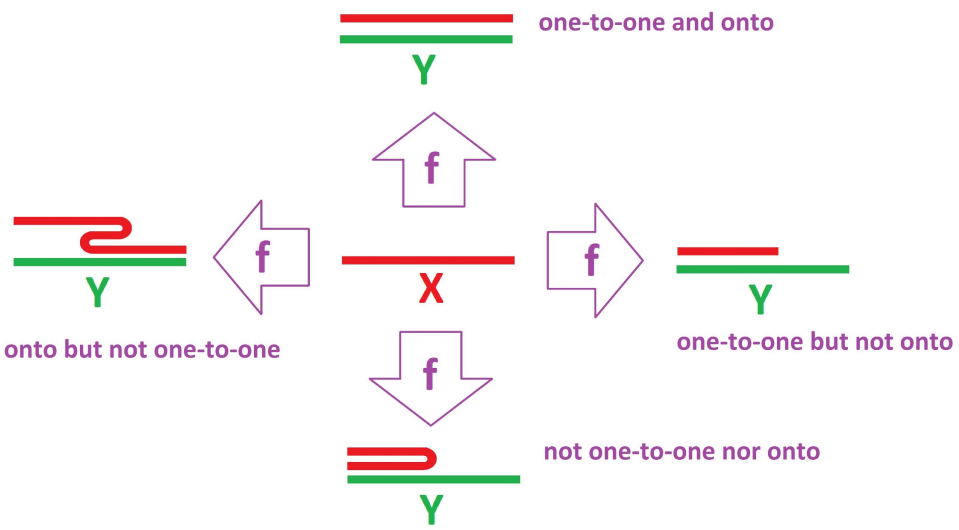
What would make a transformation to be not one-to-one? Any folding presents a visible problem:



And so does any collapsing:



So, if we imagine that the x-axis, X, is *transformed* somehow and then placed on top of the y-axis, Y, we can understand this function as a transformation. This can happen in these four basic ways:



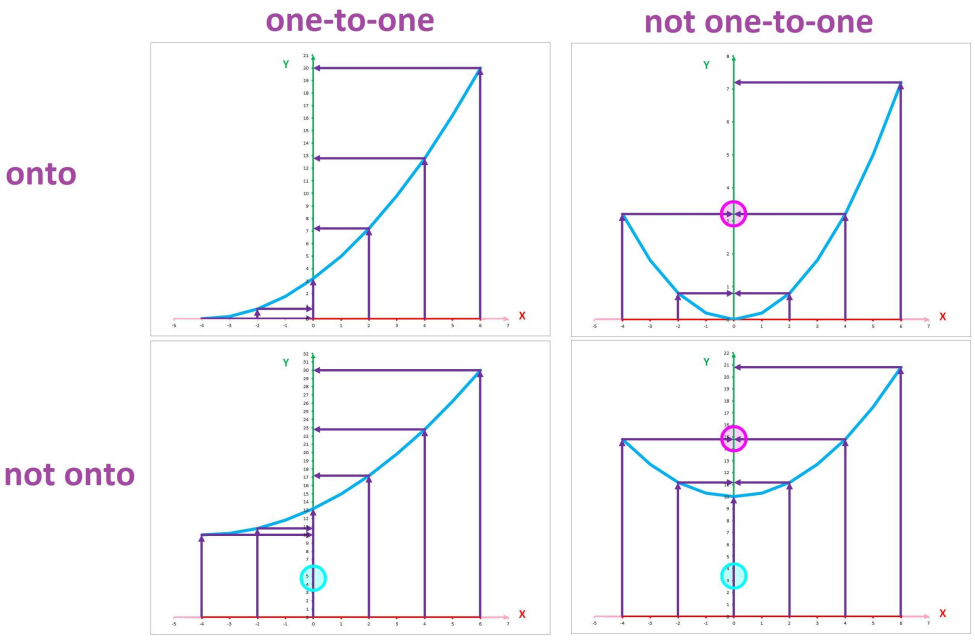
The questions we ask are these two:

- Onto: Does X cover the whole Y?
- One-to-one: Does X cover any location on Y no more than once?

Now, the graphs. The idea is to "sample" the function and see how the arrows behave.

Example 3.4.4: one-to-one, onto from graph

Let's start with this case: The domain is the reals $X = \mathbf{R} = (-\infty, +\infty)$, while the codomain is the non-negative reals, $Y = [0, \infty)$. If the graph of a function $f : X \to Y$ is given, can we determine whether this function is one-to-one or onto by just examining the graph? We arrange the four options in a 2×2 table:



We notice that:

1. The arrows in the first row seem to cover the whole Y.
2. The arrows in the first column seem not to cover any location on Y twice.

We conclude that:

1. The functions in the first row are onto.
2. The functions in the first column are one-to-one.

We draw these conclusions with the understanding that they are based on the *partial* information provided by the graph. The case for *not* onto and *not* one-to-one is better; we can definitely see and mark with circles the following:

- The arrows in the second row are missing a specific value in Y.
- The arrows in the second column are hitting an identical specific value in Y.

Exercise 3.4.5

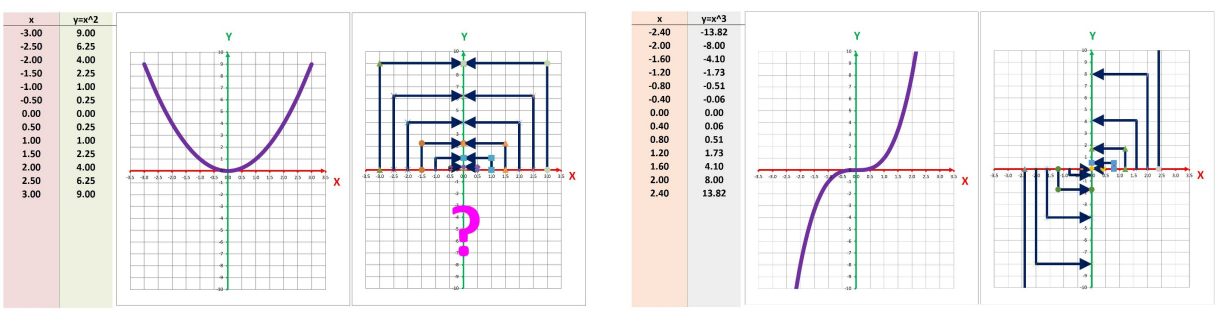
How are the conclusions in the above example affected by different choices of (a) codomains, (b) domains?

Example 3.4.6: power functions, onto

Suppose this time the domain and the codomain are the reals, $X = Y = \mathbf{R}$. Then the function $f(x) = x^2$ is not onto but $g(x) = x^3$ is. It is easy to prove the former:

$$x^2 \neq -1\,, \quad \texttt{FOR EACH } x.$$

To prove the latter will require some algebra even though the idea is clear:



Exercise 3.4.7

How can we "make" these functions onto: (a) x^2, (b) $|x|$?

What makes a difference? The graph doesn't progress below a certain line! The following is a useful observation:

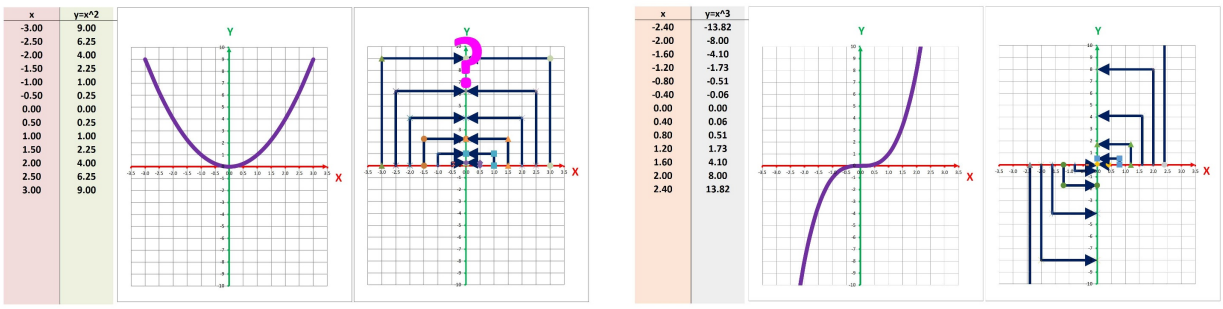
▶ A function is onto `IF AND ONLY IF` its graph and any *horizontal line* have at least one point in common.

Example 3.4.8: power functions, one-to-one

The function $f(x) = x^2$ is not one-to-one but $g(x) = x^3$ is. It is easy to prove the former:

$$1^2 = (-1)^2 = 1\,.$$

To prove the latter will require some algebra even though the idea is clear:



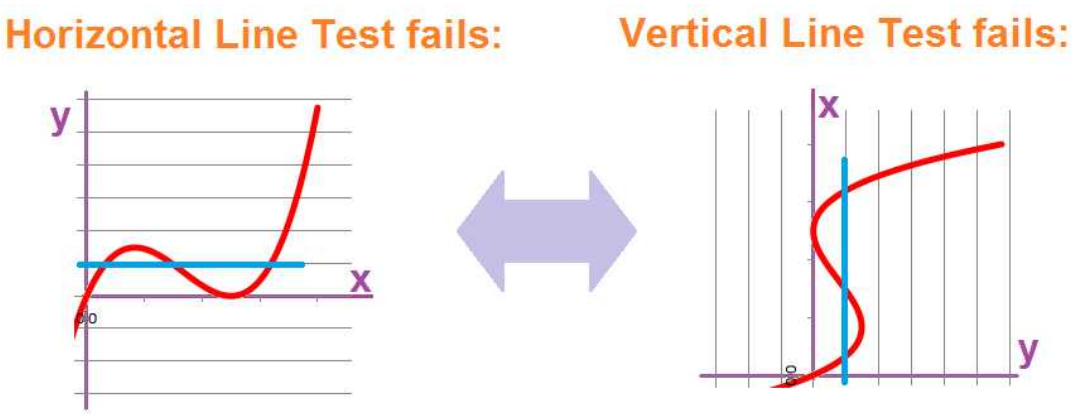
Exercise 3.4.9

How can we "make" these functions one-to-one: (a) x^2, (b) $|x|$?

What makes a difference? Two points on the graph have the same height above the x-axis! The following is a useful observation:

▶ A function is one-to-one `IF AND ONLY IF` the intersection of its graph and any *horizontal line* contains at most one point.

Notice the connection with the *Vertical Line Test for Relations* ("is this a function?") from Chapter 2:

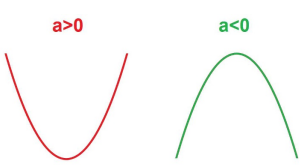


Example 3.4.10: quadratic polynomials

Quadratic polynomials,

$$F(x) = ax^2 + bx + c,\ a \neq 0\,,$$

are neither one-to-one nor onto:



The shape of the graphs of these functions will be proven in this chapter.

Exercise 3.4.11

Identify and classify the functions below:

Name	Nationality	Position	Goals
Arthur Albiston	Scotland	FB	7
Anderson	Brazil	MF	9
John Aston Jr.	England	FW	27
John Aston Sr.	England	FB	30
Gary Bailey	England	GK	0
Tommy Bamford	Wales	FW	57
Frank Barrett	Scotland	GK	0
Frank Barson	England	HB	4
Fabien Barthez	France	GK	0
Bobby Beale	England	GK	0
David Beckham	England	MF	85
Alex Bell	Scotland	HB	10
Ray Bennion	Wales	HB	3
Dimitar Berbatov	Bulgaria	FW	56
Henning Berg	Norway	DF	3
Johnny Berry	England	FW	45
George Best	Northern Ireland	FW	179

Exercise 3.4.12

A function $y = f(x)$ is given below by a list of some of its values. Add missing values in such a way that the function is one-to-one.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		4	5		2	

Exercise 3.4.13

What codomain of the function given above should we assume to assure that it is onto?

We summarize the results in the following important theorem.

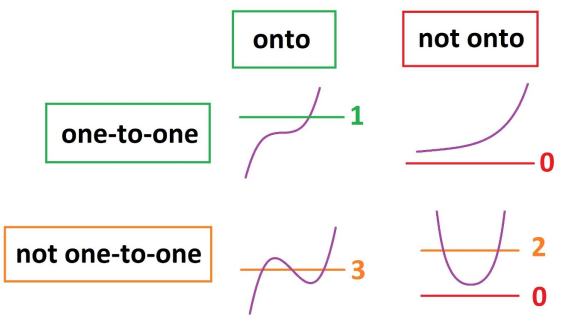
Theorem 3.4.14: Horizontal Line Test

1. *A function is onto if and only if its graph and any horizontal line have at least one point in common.*
2. *A function is one-to-one if and only if its graph and any horizontal line have at most one point in common.*

Exercise 3.4.15

Break either part of the theorem into a statement and its converse.

In light of the theorem, we redo the table presented earlier by counting the number of intersections of horizontal lines with the graph:

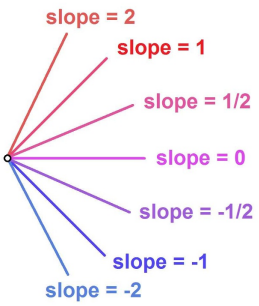


In each cell of the table, we see an example of how – in the simplest possible way – these two conditions can be violated, for a numerical function.

Exercise 3.4.16

Illustrate each of the four possibilities above with a choice of a function given by a list of values.

Linear functions are easy to characterize this way:



Indeed, as we know from geometry, two straight lines have exactly one intersection unless they are parallel; therefore the *Horizontal Line Test* is passed by all linear functions except the ones with zero slope. Those are constant functions.

However, we need to prove these facts algebraically.

Example 3.4.17: linear polynomials

Suppose we have a linear function with slope 3 and y-intercept 2:

$$F(x) = 3x + 2\,.$$

Is it one-to-one? Suppose we have two inputs x_1 and x_2. Can their outputs be equal under F? Let's try: Suppose $F(x_1) = F(x_2)$. We substitute and obtain the following:

$$3x_1 + 2 = 3x_2 + 2\,.$$

Canceling 2 produces the following:

$$3x_1 = 3x_2\,.$$

Finally, we divide by 3:

$$x_1 = x_2\,.$$

No, the outputs are equal only when the inputs are!

Is it onto? Suppose we have an input y. Is there an x taken to y under F? We just need to solve the equation $F(x) = y$ for x, for each y. We have an equation:

$$3x + 2 = y\,,$$

with an unspecified y. No matter what y is though, we subtract 2 and then divide by 3, producing:

$$x = \frac{y-2}{3}\,.$$

Yes, there is such an x, for each y!

Theorem 3.4.18: Linear Functions, One-to-one Onto

A linear function with slope m and y-intercept b,

$$F(x) = mx + b\,,$$

is both one-to-one and onto as long as $m \neq 0$.

Proof.

One-to-one. Suppose we have two inputs x_1 and x_2. Then:

$$F(x_1) = F(x_2) \implies mx_1 + b = mx_2 + b \implies mx_1 = mx_2 \implies x_1 = x_2 .$$

It's the same input.

Onto. Suppose we have an output y. We solve the equation $F(x) = y$ for x:

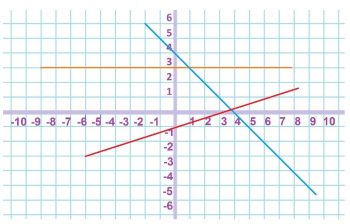
$$mx + b = y \implies x = \frac{y - b}{m} .$$

There is such an x.

Exercise 3.4.19

Break the theorem into a statement and its converse.

The picture confirms the conclusion:



Example 3.4.20: reciprocal

Let's prove algebraically that $f(x) = 1/x$ is one-to-one but not onto. There is an equation for each y with respect to x:

$$y = 1/x .$$

We solve each for x:

$$y = 1/x \implies x = 1/y .$$

A single solution for each y, hence the function is one-to-one. But there is no solution when $y = 0$, hence the function isn't onto. Of course, the answer depends on your choice of the codomain: $Y = \{y : y \neq 0\}$ makes $f : \mathbf{R} \to Y$ onto.

Exercise 3.4.21

Prove algebraically that $f(x) = 1/x^2$ is neither one-to-one nor not onto. Show how the answer changes with your choice of the codomain.

Exercise 3.4.22

Classify the function $f(x) = x^3 - x$ according to these two definitions. Prove algebraically.

Theorem 3.4.23: One-to-one and Onto vs. Image

1. *A function $F : X \to Y$ is onto if and only if its image is the whole codomain, Y.*
2. *A function $F : X \to Y$ is one-to-one if and only if the preimage of every element of the codomain Y is a single element of the domain, X, or it's empty.*

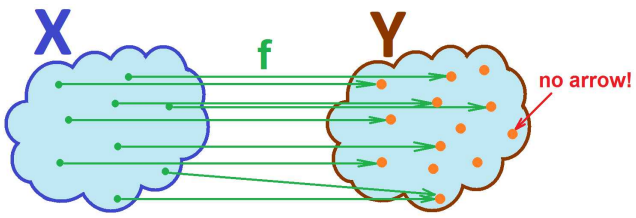
Exercise 3.4.24

Break either part of the theorem into a statement and its converse.

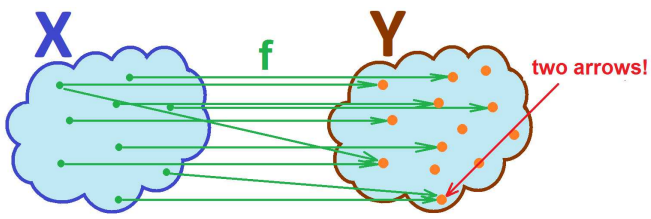
Exercise 3.4.25

Prove the theorem.

To summarize, the restrictions in these two definitions can be violated when there are too few or too many arrows arriving to a given y. These violations are seen in the codomain. This one is *not onto*:

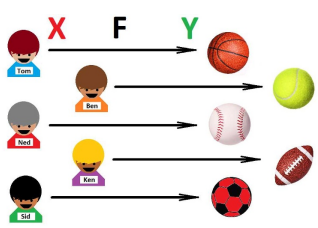


That one is *not one-to-one*:



Example 3.4.26: both one-to-one and onto

What functions are the most "regular"? The ones that are both one-to-one and onto, sometimes called *bijections*:



The function may look plain but it has an important property: the boys and the balls can be used as substitutes of each other! For example, you don't have to remember the name of every boy but just say "the one that plays basketball" to identify Tom without a chance of confusion.

Exercise 3.4.27

What are the smallest set X and the smallest set Y for which a function $F : X \to Y$ can be: (a) not one-to-one, (b) not onto?

Exercise 3.4.28

Sketch the graph of the function f given by its list of values below. Is it one-to-one?

x	1	2	3	4	5
$y = f(x)$	1	2	0	3	1

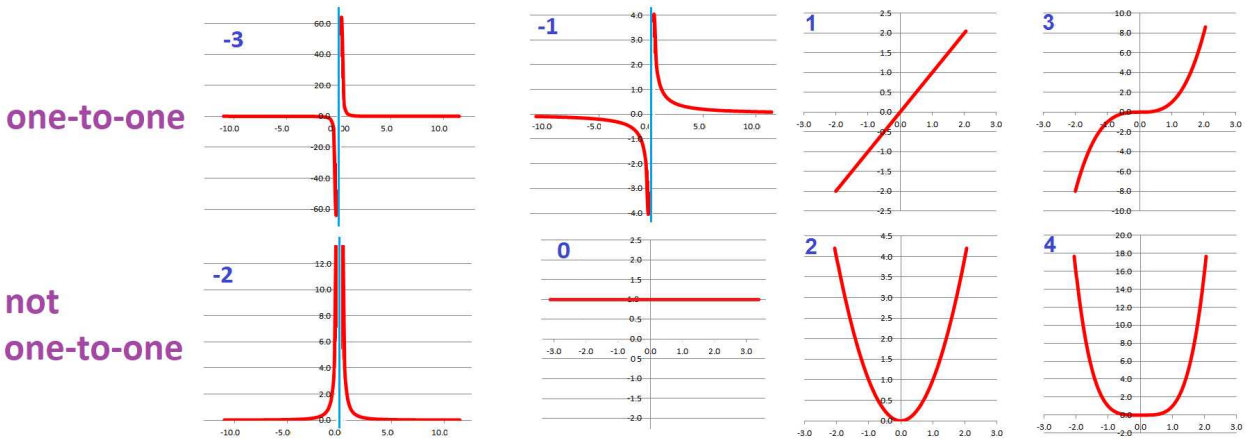
Make up your own functions by providing their lists of values and test the two definitions. Repeat.

Example 3.4.29: power functions, symmetries

The power functions,

$$y = x^n, \ n = ..., -3, -2, -1, 0, 1, 2, 3, ...,$$

can now be classified:



Exercise 3.4.30

Which ones are onto?

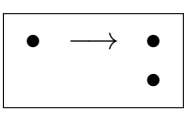
Exercise 3.4.31

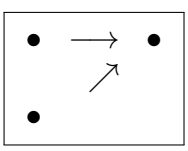
Prove that a one-to-one function can't have mirror symmetry.

Exercise 3.4.32

What kind of function is "many-to-one"? What about "one-to-many"?

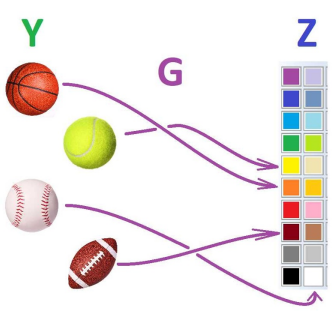
In summary, we present the two simplest ways the two conditions can be *violated*:

Not onto: 

Not 1-1: 

3.5. Compositions of functions

Back to our boys-and-balls example, let's note the *colors of the balls*. This has nothing to do with the boys, and it creates a new function:

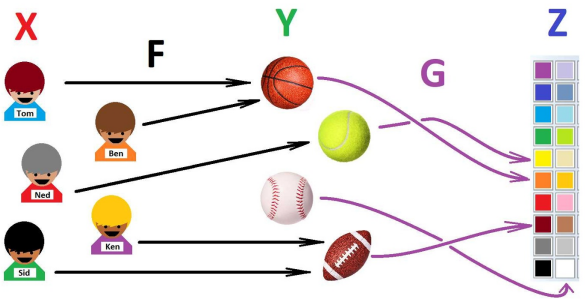


It is a function $G : Y \to Z$ from the set of all balls to the new set Z of the main colors.

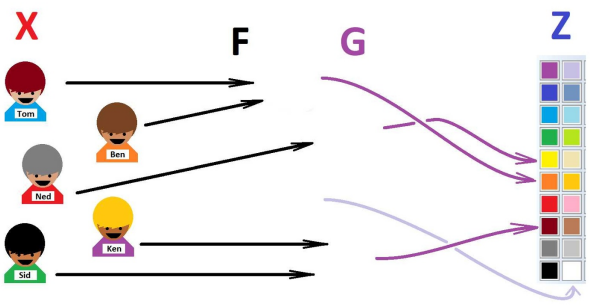
Exercise 3.5.1

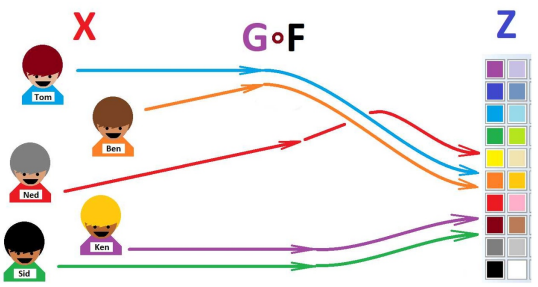
Is the new function one-to-one or onto?

We know the boys' preferences in balls, but does it entail any *preferences in colors*? In a sense. We can just combine the new function with the old:

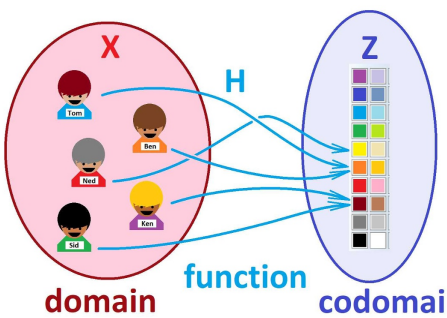


If we start with a boy on the left, we can continue with the arrows all the way to the right. This way, we will know the color of the ball the boy is playing with:





This is a new function, say $H : X \to Z$, from the set of boys X to the set Z of the main colors:



Exercise 3.5.2

Is the new function one-to-one or onto?

In general, this is the setup:

$$X \xrightarrow{F} Y \xrightarrow{G} Z$$

The following concept is very important.

Definition 3.5.3: composition of functions

Suppose we have two functions (with the codomain of the former matching the domain of the latter):

$$F : X \to Y \text{ and } G : Y \to Z\,.$$

Then their *composition* is the function (from the domain of the former to the codomain of the latter)

$$H : X \to Z\,,$$

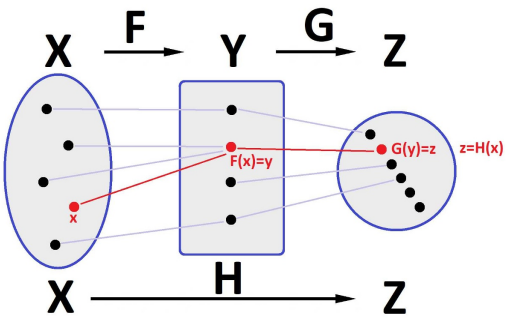
which is computed for every x in X according to the following two-step procedure:

$$x \mapsto F(x) = y \mapsto G(y) = z$$

It is denoted by

$$G \circ F$$

We just follow from X along the arrows of F to Y and then along the arrows of G to Z:



Warning!

The requirement that "the codomain of the former is the domain of the latter" could be replaced with "the domain *contains* the codomain".

In other words, the new function is evaluated by the *substitution formula*:

$$z = H(x) = G(F(x))$$

This is the "deconstruction" of the notation:

Composition of functions

$$\underset{\substack{\uparrow\\ \text{name of the new function}}}{(G \circ F)(x)} = \underset{\substack{\text{names of the second and first functions}\\ \downarrow \quad \downarrow\\ G\Big(\quad F(x) \quad\Big)\\ \uparrow\ \uparrow\ \uparrow\\ \text{substitution}}}{}$$

The name of the function on left reads "G composition F". The computation on right reads "G of F of x".

It is not that the operations are read from right to left, but rather *from inside out*!

If we represent the two functions as *black boxes*, we can wire them together:

$$\begin{array}{ccccccccc} \text{input} & & \text{function} & & \text{output} & & & & \\ x & \to & \boxed{F} & \to & y & & & & \\ & & & & \downarrow & & & & \\ & & & & \text{input} & & \text{function} & & \text{output} \\ & & & & y & \to & \boxed{G} & \to & z \end{array}$$

Thus, we use the output of the former as the input of the latter.

To make it clear that Y is no longer a part of the picture, we can also visualize the composition as follows:

$$\begin{array}{ccc} X & \xrightarrow{F} & Y \\ & \searrow^{H} & \downarrow G \\ & & Z \end{array}$$

The meaning of the diagram is as follows: Whether we follow the F-then-G route or the direct H route, the results will be the same.

If we think of functions as *lists of instructions*, we just attach the list of the latter at the bottom of the list of the former. In other words, here is the list of $G \circ F$:

- Step 1: Do F.
- Step 2: Do G.

They are executed *consecutively*; you can't start with the second until you are done with the first.

Let's test this idea on our example. We take the two *lists of values* and then cross-reference them (from left to right):

$$\boxed{\begin{array}{lcl} F(\text{ Tom }) & = & \text{basketball} \\ F(\text{ Ned }) & = & \text{tennis} \\ F(\text{ Ben }) & = & \text{basketball} \\ F(\text{ Ken }) & = & \text{football} \\ F(\text{ Sid }) & = & \text{football} \end{array}} \quad , \quad \boxed{\begin{array}{lcl} G(\text{ basketball }) & = & \text{orange} \\ G(\text{ tennis }) & = & \text{yellow} \\ G(\text{ football }) & = & \text{brown} \\ G(\text{ baseball }) & = & \text{white} \end{array}} \quad \longrightarrow \quad \boxed{\begin{array}{lcl} H(\text{ Tom }) & = & ? \\ H(\text{ Ned }) & = & ? \\ H(\text{ Ben }) & = & ? \\ H(\text{ Ken }) & = & ? \\ H(\text{ Sid }) & = & ? \end{array}}$$

We ignore any alignment between the two lists. We take the first entry in the second list,

$$G(\text{ basketball }) = \text{ orange},$$

and replace "basketball", according to the first entry of the first list, with $F(\text{ Tom })$. This is the result:

$$G\Big(\ F(\text{ Tom })\ \Big) = \text{ orange}.$$

Therefore,

$$H(\text{ Tom }) = \text{ orange}.$$

This is the first entry in the new list.

Exercise 3.5.4

Finish the list.

Once again, algebraically, composition is nothing but *substitution*!

Unfortunately, the domains of numerical functions are often infinite, and it is impossible to just follow the arrows. We deal with formulas. Fortunately, substitution is just as simple for numerical functions.

Example 3.5.5: composition of numerical functions

This time, we substitute one formula into another. For example, consider these two functions:

$$\boxed{X \xrightarrow{x^2=y} Y \xrightarrow{y^3=z} Z}$$

Then, $y = x^2$ is substituted into $z = y^3$ resulting in:

$$z = \left(x^2\right)^3.$$

It is the same function, and it is computed by the same *two* steps! A simplification might make the extra work worthwhile:

$$z = x^6\,.$$

Example 3.5.6: substitution of numerical functions

The idea of how the substitution is executed is the same as in Chapter 2: Insert the input value in all of these boxes. Suppose this function on the left is understood and evaluated via the diagram on the right:

$$f\,(y) = \frac{2y^2 - 3y + 7}{y^3 + 2y + 1}, \quad f\,(\square) = \frac{2\square^2 - 3\square + 7}{\square^3 + 2\square + 1}\,.$$

Previously, we did the substitution $y = 3$ by inserting 3 in each of these windows:

$$f\left(\boxed{3}\right) = \frac{2\boxed{3}^2 - 3\boxed{3} + 7}{\boxed{3}^3 + 2\boxed{3} + 1}.$$

This time, let's insert $\sin x$, or, better, $(\sin x)$. This is the result of the substitution $y = \sin x$:

$$f\boxed{(\sin x)} = \frac{2\boxed{(\sin x)}^2 - 3\boxed{(\sin x)} + 7}{\boxed{(\sin x)}^3 + 2\boxed{(\sin x)} + 1}.$$

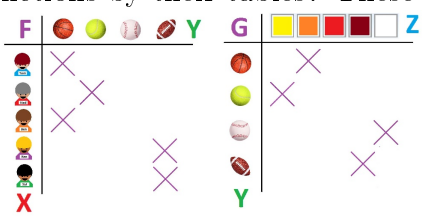
Then, we have

$$f(\sin x) = \frac{2(\sin x)^2 - 3(\sin x) + 7}{(\sin x)^3 + 2(\sin x) + 1}.$$

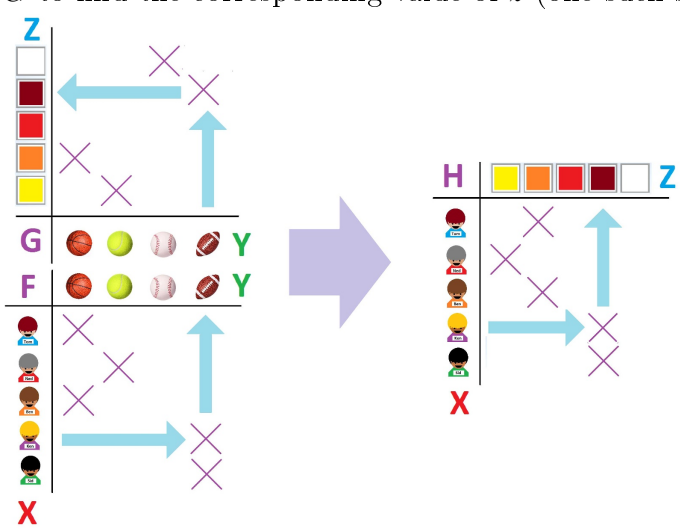
Note that if you don't know what the sine function (Chapter 4) does, it makes no difference! The only thing that matters is that we know that this is a *function*.

Example 3.5.7: composition from tables

Next, what about representing functions by their tables? These are the tables of F and G above:



How do we combine them to find H? The alignment we might see is meaningless. We need to align what the two have in common, Y. Then we start with an x in X, use F to find the corresponding value of y in Y, then use G to find the corresponding value of z (one such step is illustrated below):



Example 3.5.8: composition from graphs

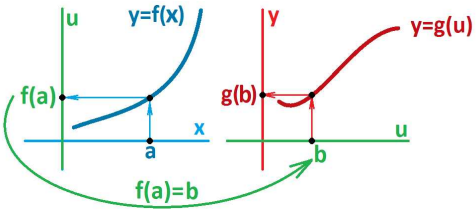
What if we have only their graphs? Suppose we have the two graphs of

$$u = f(x) \quad \text{and} \quad y = g(u)\,,$$

side by side and we need to find the composition, $g \circ f$. Let's take a *single* value:

$$d = g(f(a))\,.$$

Then, we use the first graph to find $c = f(a)$ on the vertical axis, travel all the way to the horizontal axis of the second, find the corresponding c on it, and finally find $d = g(c)$ on the vertical axis.



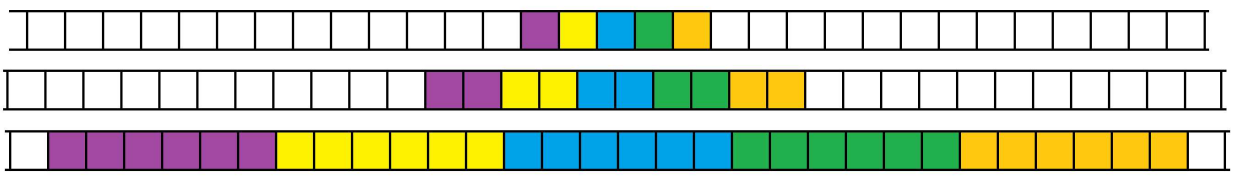
Graphs don't do a very good job of visualizing compositions...

Example 3.5.9: composition of transformations

What if we, instead, think of numerical functions as transformations? For example:

- The first transformation is a stretch by a factor of 2.
- The second transformation is a stretch by a factor of 3.

We visualize the first by doubling every square and the second by tripling:

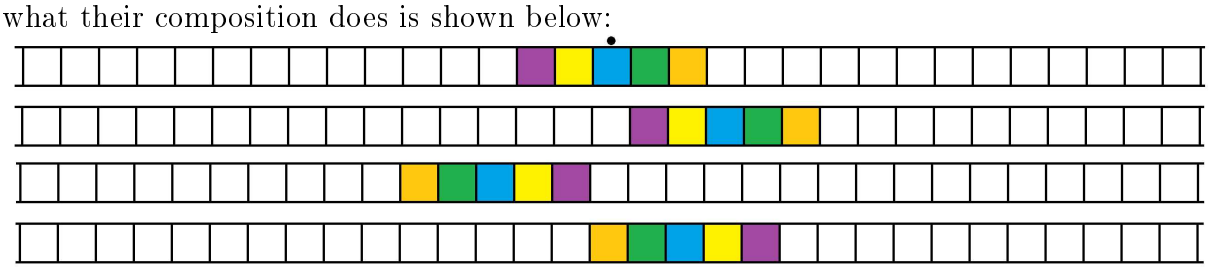


As a result there are *six* squares in the last row for each square in the first. Indeed, the composition of the two transformations is a stretch by a factor of $3 \cdot 2 = 6$.

So, we just carry out two transformations in a row. Consider this also:

- The first transformation is a shift (right) by a 3.
- The second transformation is a flip.
- The third transformation is a shift (right) by a 5.

Then what their composition does is shown below:



Exercise 3.5.10

Illustrate, as above, the composition of: a shift left by 5, a stretch by 2, and a flip.

Exercise 3.5.11

Illustrate the composition of the two transformations shown below:



If we think of functions as *lists of instructions* (with no forks), then each of them is already a composition! The steps on the list are the functions the composition of which creates the function. For example,

▶ F: Add 3, multiply by -2, subtract 1.

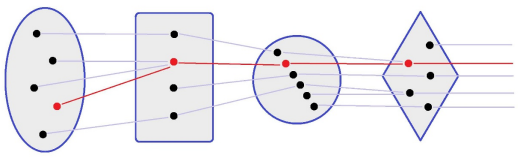
This is called a *decomposition* of F. If, furthermore, there is another function, say,

▶ G: Subtract 2, apply sin.

Now, we just add the latter list to the bottom of the former:

▶ $G \circ F$: Add 3, multiply by -2, subtract 1, subtract 2, apply sin.

Of course, we can have compositions of many functions in a row as long the output of each function matches the input of the next:



It's as if the first function gives us direction to a destination, and at that destination, we receive the directions to our next destination where we get further directions, and so on... like a treasure hunt.

We represent functions as *black boxes* that process the input and produce the output:

$$\begin{array}{ccccccccccc} \text{input} & & \text{function} & & \text{output} & & & & & & \\ x & \to & \boxed{f} & \to & y & & & & & & \\ & & & & \text{input} & & \text{function} & & \text{output} & & \\ & & & & y & \to & \boxed{g} & \to & z & & \\ & & & & & & & & \text{input} & \text{function} & \text{output} \\ & & & & & & & & z & \to \boxed{h} \to & u \end{array}$$

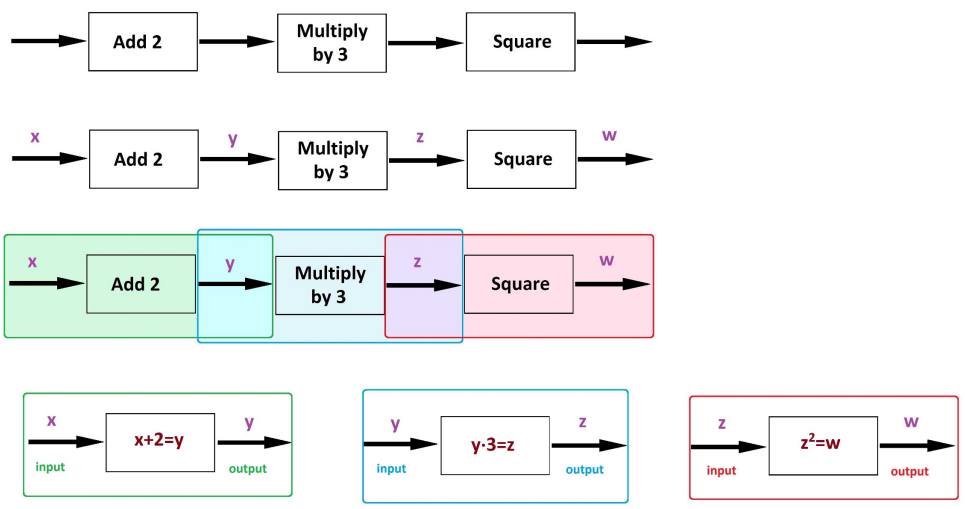
Because of the match, we can carry over the output to the next line – as the input of the next function.

This *chain of events* can be as long as we like:

$$x_1 \to \boxed{f_1} \to x_2 \to \boxed{f_2} \to x_3 \to \boxed{f_3} \to x_4 \to \boxed{f_4} \to x_5 \to \ldots$$

Example 3.5.12: decomposition of flowchart

Such a decomposition will allow us to study the function one piece at a time:



Exercise 3.5.13

Represent the following function by a single formula:

$$x \to \boxed{\text{multiply by 2}} \to y \to \boxed{\text{add 5}} \to z \to \boxed{\text{divide by 3}} \to u$$

Exercise 3.5.14

Represent the function $h(x) = (x-1)^2 + (x-1)^3$ as the composition $g \circ f$ of two functions $y = f(x)$ and $z = g(y)$.

Example 3.5.15: composition with spreadsheet, formulas

This is how the composition of several functions represented by *formulas* is computed with a spreadsheet. We start with just a list of numbers in the first column. Then we produce the values in the next column one row at a time. How? We input a formula in the next column with a reference to the last one. For example, we have in the second and third columns, respectively:

`=RC[-1]*2` `=RC[-1]+5`

referring to the cell to its left. Each of the two consecutive columns is a list of values of a function (left):

x	y	z
0	0	5
1	2	7
2	4	9
3	6	11
4	8	13
5	10	15
6	12	17
7	14	19
8	16	21
9	18	23
10	20	25
11	22	27
12	24	29
13	26	31
14	28	33
15	30	35
16	32	37
17	34	39
18	36	41
19	38	43

x	z
0	5
1	7
2	9
3	11
4	13
5	15
6	17
7	19
8	21
9	23
10	25
11	27
12	29
13	31
14	33
15	35
16	37
17	39
18	41
19	43

If we hide the middle column (right), we have the list of values of the composition. We can have as many intermediate columns as we like.

Example 3.5.16: functions given by lists, cross-referencing

What if the functions are given by nothing but their lists of values? Then we need to find a match for the output of the first function among the inputs of the second. Given the tables of values of f, g, find the table of values of $g \circ f$:

x	$y = f(x)$
0	1
1	0
2	2
3	4
4	2

followed by

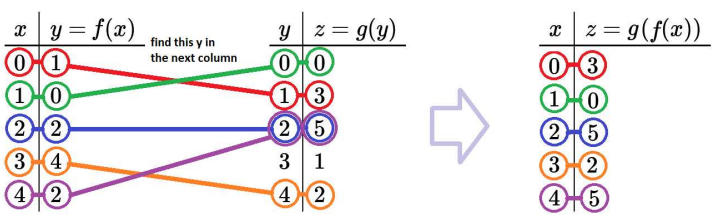
y	$z = g(y)$
0	0
1	3
2	5
3	1
4	2

is

x	$z = g(f(x))$
0	?
1	?
2	?
3	?
4	?

.

We need to fill the second column of the last table. First, we need to match the outputs of f with the inputs of g, as follows:



Alternatively, we re-arrange the rows of g according to the values of y and then remove the y-columns:

x	y	y	z
0	1	1	3
1	0	0	0
2	2	2	5
3	4	4	2
4	2	2	5

$\longrightarrow$

x	z
0	3
1	0
2	5
3	2
4	5

Exercise 3.5.17

Given the tables of values of f, g below, find the table of values of $g \circ f$:

x	$y = f(x)$
0	0
1	2
2	3
3	0
4	1

y	$z = g(y)$
0	4
1	4
2	0
3	1

Example 3.5.18: composition with spreadsheet, cross-referencing

For this task, we have to use the search feature of the spreadsheet:

R3C8 =VLOOKUP(RC[-6],R3C[-4]:R18C[-3],2)

	1	2	3	4	5	6	7	8
1		f			g			
2	t	x=f(t)		x	y=g(x)		t	y=g(f(t))
3	0	1		1	1		0	1
4	1	2		2	4		1	4
5	2	3		3	9		2	9
6	3	1		4	16		3	1
7	4	4		5	25		4	16
8	5	3		6	36		5	9
9	6	2		7	49		6	4
10	7	1		8	64		7	1
11	8	1		9	81		8	1
12	9	1		10	100		9	1
13	10	1		11	121		10	1
14	11	4		12	144		11	16
15	12	5		13	169		12	25
16	13	6		14	196		13	36
17	14	11		15	225		14	121
18	15	1		16	256		15	1

For example, the search may be executed with a "look-up" function:

```
=VLOOKUP(RC[-6],R3C[-4]:R18C[-3],2)
```

Example 3.5.19: two functions, two compositions

Consider these functions:

$$f(x) = x^2\,, \quad g(y) = y + 1\,.$$

Both $f \circ g$ and $g \circ f$ make sense. We just need to indicate which input, x or y, will as the output of the other function. The former:

$$f(x) = x^2\,, \quad x = g(y) = y + 1\,.$$

We now know what to substitute:

$$(f \circ g)(y) = f(g(y)) = (y + 1)^2\,.$$

The latter:

$$y = f(x) = x^2\,, \quad g(y) = y + 1\,.$$

We now know what to substitute:

$$(g \circ f)(y) = g(f(x)) = (x^2) + 1\,.$$

Example 3.5.20: unspecified substitution

You may encounter problems stated as follows:

▶ Find the composition of the functions:

$$f(x) = \frac{x}{1+x}\,, \quad g(x) = \sin x\,.$$

Since both functions have the *same* input variable, it is unclear how to substitute! Is it f then g, or g then f? Let's sort this out and do both.

The flowchart of the composition $f \circ g$ is the following:

$$x \;\to\; \boxed{g} \;\to\; y \;\to\; \boxed{f} \;\to\; z$$

This means that in order to make the output of the first to match the input of the second, we have to *rename the variables*:

$$x \;\to\; \boxed{\sin x} \;\to\; y \;\to\; \boxed{\frac{y}{1+y}} \;\to\; z$$

This is the new version of the problem:

▶ Find the composition, $f \circ g$, of the functions:

$$f(y) = \frac{y}{1+y}\,, \quad y = g(x) = \sin x\,.$$

Now we substitute

$$y = \sin x \;\text{ into }\; f(y) = \frac{y}{1+y}\,,$$

as follows:

$$\begin{aligned}(f \circ g)(x) \;&= f(y) \;&&= f(g(x))\\ &= \frac{y}{1+y} \;&&= \frac{\sin x}{1+\sin x}\,.\end{aligned}$$

Now, the reverse. The flowchart of the composition $g \circ f$:

$$x \;\to\; \boxed{f} \;\to\; y \;\to\; \boxed{g} \;\to\; z$$

Once again, we have to rename the variables:

$$x \;\to\; \boxed{\frac{x}{1+x}} \;\to\; y \;\to\; \boxed{\sin y} \;\to\; z$$

Now we substitute

$$y = \frac{x}{1+x} \;\text{ into }\; g(y) = \sin y\,,$$

as follows:

$$\begin{aligned}(g \circ f)(x) \;&= g(y) \;&&= g(f(x))\\ &= \sin y \;&&= \sin\left(\frac{x}{1+x}\right)\,.\end{aligned}$$

Exercise 3.5.21

Represent the function $h(x) = 2(\sin x)^3 + 4\sin x + 5$ as the composition of two functions one of which is trigonometric. Hint: It doesn't matter what sin does.

Exercise 3.5.22

(a) Represent the function $h(x) = \sqrt{x-1}$ as the composition of two functions. (b) Represent the function $k(t) = \sqrt{t^2-1}$ as the composition of three functions. (c) Represent the function $p(t) = \sin\left(\sqrt{t^2-1}\right)$ as the composition of four functions.

Exercise 3.5.23

Functions $y = f(x)$ and $u = g(y)$ are given below by tables of some of their values. Present the composition $u = h(x)$ of these functions by a similar table:

x	0	1	2	3	4
$y = f(x)$	1	1	2	0	2

y	0	1	2	3	4
$u = g(y)$	3	1	2	1	0

Exercise 3.5.24

Represent the composition of these two functions: $f(x) = 1/x$ and $g(y) = \dfrac{y}{y^2 - 3}$, as a single function h of variable x. Don't simplify.

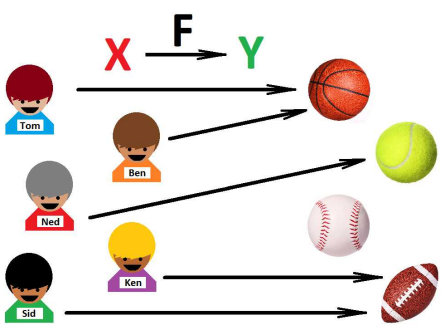
Exercise 3.5.25

Function $y = f(x)$ is given below by a list of its values. Is the function one-to-one?

x	0	1	2	3	4
$y = f(x)$	0	1	2	1	2

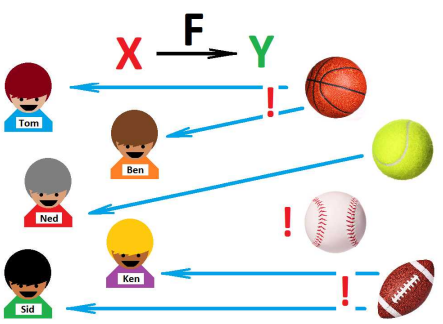
3.6. The inverse of a function

Our running example of a function answers the question: *Which ball is this boy playing with?*



Let's turn this around: *Which boy is playing with this ball?*

The solution would seem to be a simple reversal of the arrows:



We can see that, even though the latter question is asked about the same situation as the former, it cannot be answered in a positive manner! Indeed:

- There are two boys playing with the basketball – two answers.

- There is no boy playing with the baseball – no answer.

This means that there is no function this time!

Question: *Under what circumstances would such a reversal of arrows make sense?*

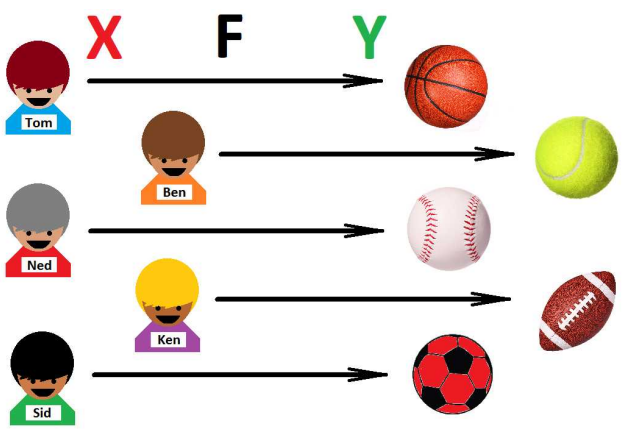
All functions *from* X *to* Y are also relations *between* X and Y. However, not every relation is a function – either from X to Y or from Y to X. The reasons are the same: too many or too few arrows starting at an element of the domain set.

An especially important question is: Can we reverse the arrows of a function so that that the same relation is now seen as a new, "inverse", function? If the answer is No, can we see the reason by looking at the *original* function?

- First, some y's in Y have no corresponding x's in X. In other words, the function isn't onto!
- Second, some y's in Y have two or more corresponding x's in X. In other words, the function isn't one-to-one!

So, the original function lacks either of the two types of regularity for this to be possible.

What kind of function would make this possible? A function that is *both one-to-one and onto*:



We have added an extra ball (soccer) and have re-drown the arrows; there is exactly one arrow for each ball. This is a *very simple*, almost uninteresting, kind of function: each boy holds a single ball, and every ball is held by a single boy.

As a result, the only difference between the two sets is in the names:

boys	balls
Tom	basketball
Ben	tennis
Ned	baseball
Ken	football
Sid	soccer

Indeed, if you don't remember Tom's name, you just say "the boy who plays with the basketball". Or, if you don't remember what that red ball is for, you just say "the game Sid plays". There is no ambiguity in this substitution.

The following is very important.

Definition 3.6.1: inverse of function

Suppose $F : X \to Y$ is a function. A function $G : Y \to X$ is called an *inverse function* of F when, for all x and y, we have the following match:

$$F(x) = y \text{ if and only if } G(y) = x$$

In other words, we have:

$$x \xrightarrow{F} y \iff y \xrightarrow{G} x$$

We will rely on this important fact.

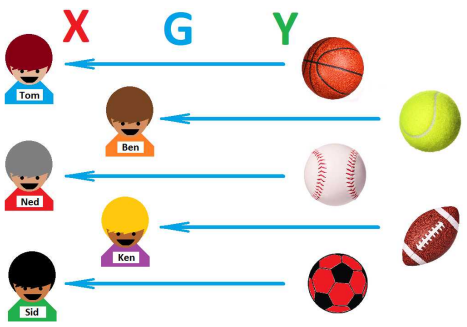
Theorem 3.6.2: One-to-one Onto vs. Inverse

A function is both one-to-one and onto if and only if it has an inverse.

Exercise 3.6.3

Prove the theorem.

Under this condition, the arrows can be safely reversed:



As the function is defined indirectly, we need an assurance that there is only one.

Theorem 3.6.4: Uniqueness of Inverse

There can be only one inverse for a function.

Exercise 3.6.5

Prove the theorem.

This justifies using "the inverse" from now on. The inverse of a function F is denoted as follows:

Inverse function

$$F^{-1}$$

It reads "F inverse".

Here "F" is the *name* of the old function and "F^{-1}" is the *name* of the new function with a reference to the one it came from.

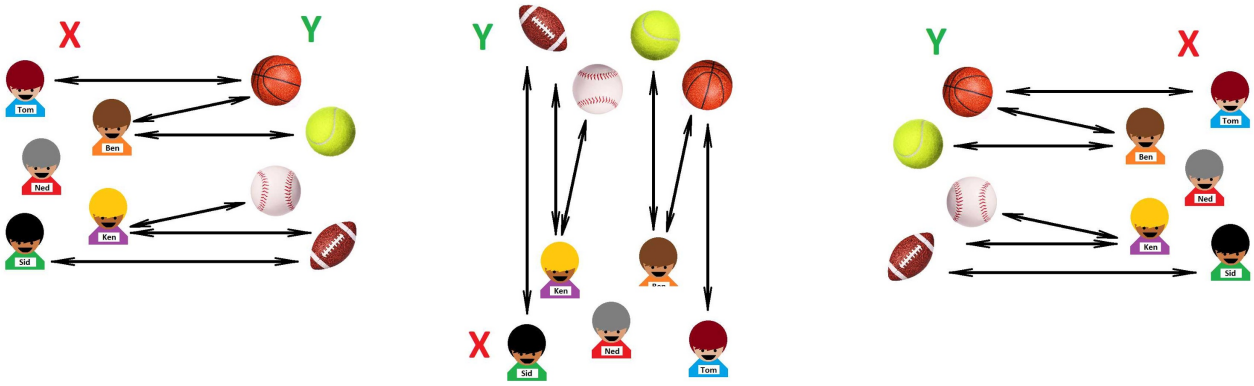
Warning!

The notation is not to be confused with the power notation for the reciprocal: $2^{-1} = \frac{1}{2}$. (Warning inside a warning: the inverse of multiplication by 2 is division by 2.)

An idea to hold on to is that a function and its inverse represent the *same relation* between sets X and Y:

- x and y are related when $y = F(x)$, or
- x and y are related when $x = F^{-1}(y)$.

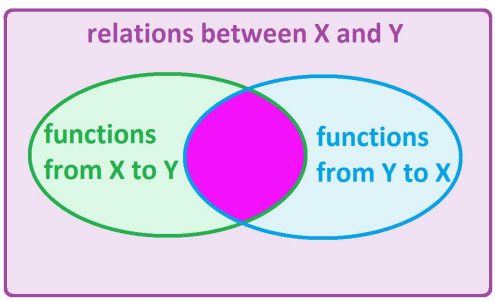
There is no preferred alignment:



The choice between F and F^{-1} is the choice of the "roles" for X and Y, input or output, domain or codomain.

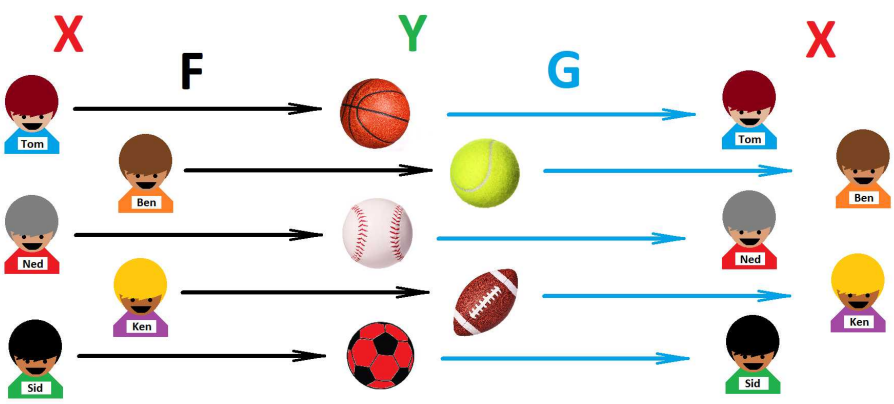
Exercise 3.6.6

Explain the picture below:



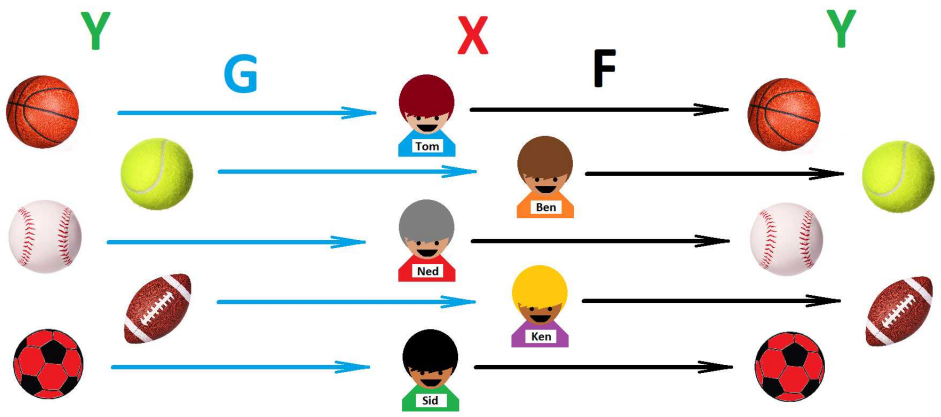
Compositions provide a different point of view on the definition of inverse.

Notice that the domain of the new function – the inverse – would have to be the codomain of the original! Their composition then makes sense:



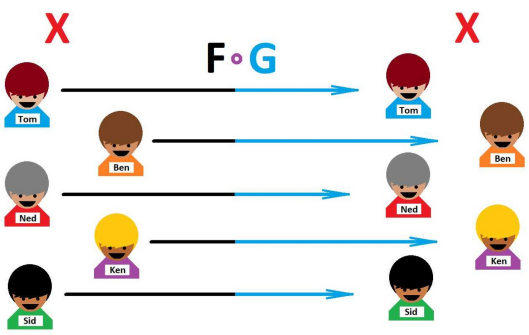
What is unusual about this composition is that we have made – through the two functions – a full circle from boys to balls and back to boys. Furthermore, with every two consecutive arrows, we arrive to our starting point, a boy.

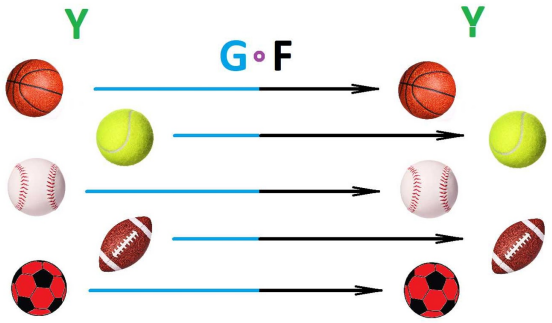
Furthermore, there is the another way (the opposite) to build a composition from these two:



We have made – through the two functions – a full circle from balls to boys and back to balls. Every time, we arrive to our starting point, a ball.

These are the two compositions next to each other:





Here is a flowchart representation of this idea:

$$x \to \boxed{F} \to y \to \boxed{G} \to x \quad \text{... same } x!$$
$$y \to \boxed{G} \to x \to \boxed{F} \to y \quad \text{... same } y!$$

We feed the output of F into G, and vice versa. The following is crucial.

Theorem 3.6.7: Inverse via Compositions

Suppose $F : X \to Y$ is a function that is both one-to-one and onto. Then a function $G : Y \to X$ is the inverse of F if and only if

- $G(F(x)) = x$ *FOR EACH* x, *AND*
- $F(G(y)) = y$ *FOR EACH* y.

Exercise 3.6.8

Prove the theorem.

The two identities in the theorem can also be written as follows:

$$\big(F^{-1} \circ F\big)(x) = x$$
$$\big(F \circ F^{-1}\big)(y) = y$$

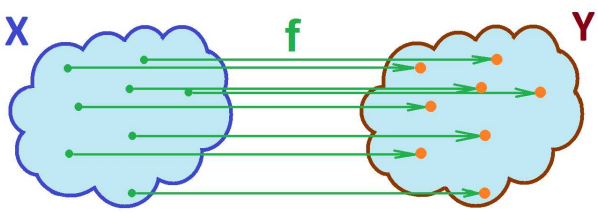
Making circles in the diagram below won't change the value of x or y:

$$\begin{array}{ccc} x & \xrightarrow{\quad F \quad} & y \\ \uparrow & & \downarrow \\ x & \xleftarrow{\quad F^{-1} \quad} & y \end{array}$$

Definition 3.6.9: invertible function

A function that is both one-to-one and onto is also called *invertible* (or a bijection).

Such a function is very easy to illustrate:



Such a function creates a correspondence between two sets that makes it look like this is the same set.

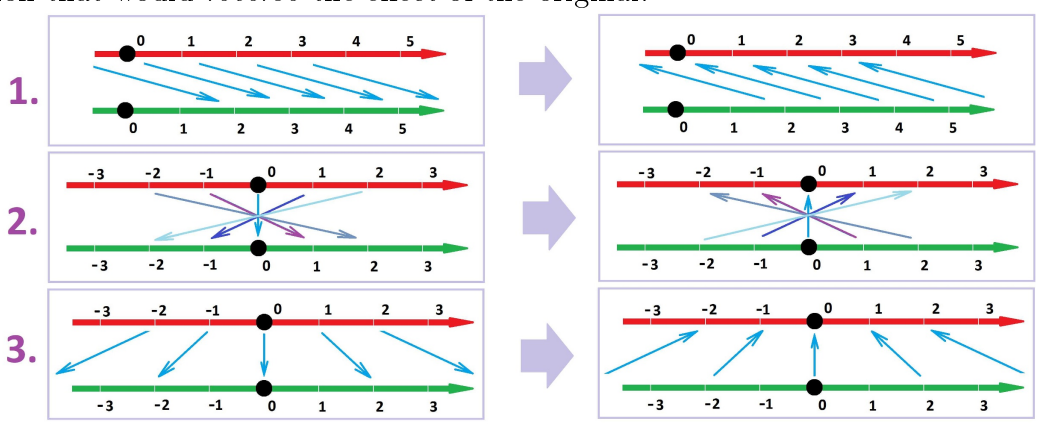
Exercise 3.6.10

Suppose A, B, C are the sets of the one-to-one, the onto, and the invertible functions, respectively. What is the relation between these sets?

Now, the *numerical functions*...

Example 3.6.11: inverse of transformation

What is the meaning of the inverse of a function when seen as a transformation of the line? It is a transformation that would *reverse* the effect of the original:



Just by examining these simple transformations, we discover the following:

1. The inverse of the shift s units to the right is the shift of s units to the left.
2. The inverse of the flip is another flip.
3. The inverse of the stretch by $k \neq 0$ is the shrink by k (i.e., stretch by $1/k$).

In fact, we realize that they *pair up*:

1. The shift s units to the right and the shift of s units to the left are the inverses of each other.
2. The flip is the inverse with itself.
3. The stretch by $k \neq 0$ and the shrink by k are the inverses of each other.

When executed consecutively (in either order), the effect is nil.

Algebraically, we have these pairs of functions:

	$f = g^{-1}$	vs.	$g = f^{-1}$
1. shift	$y = x + s$		$x = y - s$
2. flip	$y = -x$		$x = -y$
3. stretch	$y = x \cdot k$		$x = y/k$

What about the fold? It can't be undone since any two points that are brought together become

indistinguishable. Any further transformations will produce the same output:

$$\begin{array}{ccccc} x=a & & & & \\ & \searrow^{F} & & & \\ & & y & \xrightarrow{G?} & x \\ & \nearrow_{F} & & & \\ x=b & & & & \end{array}$$

This function isn't one-to-one! And neither is the collapse.

Example 3.6.12: inverse of list

How do we find the inverse of a function given by its *list of values*:

$$f = \left|\begin{array}{c|c} x & y=f(x) \\ \hline 0 & 1 \\ 1 & 0 \\ 2 & 2 \\ 3 & 1 \\ 4 & 3 \\ \dots & \dots \end{array}\right| = \left|\begin{array}{ccc} x & \to & y \\ \hline 0 & \to & 1 \\ 1 & \to & 0 \\ 2 & \to & 2 \\ 3 & \to & 1 \\ 4 & \to & 3 \\ \dots & \dots & \dots \end{array}\right|$$

The table is understood as if there are arrows going horizontally left to right. That is why "reversing the arrows" means interchanging the columns:

$$f^{-1} = \left|\begin{array}{ccc} x & \to & y \\ \hline 0 & \leftarrow & 1 \\ 1 & \leftarrow & 0 \\ 2 & \leftarrow & 2 \\ 3 & \leftarrow & 1 \\ 4 & \leftarrow & 3 \\ \dots & \dots & \dots \end{array}\right| = \left|\begin{array}{ccc} y & \to & x \\ \hline 1 & \to & 0 \\ 0 & \to & 1 \\ 2 & \to & 2 \\ 1 & \to & 3 \\ 3 & \to & 4 \\ \dots & \dots & \dots \end{array}\right| = \left|\begin{array}{ccc} y & \to & x \\ \hline 0 & \to & 1 \\ 1 & \to & 0 \\ 1 & \to & 3 \\ 2 & \to & 2 \\ 3 & \to & 4 \\ \dots & \dots & \dots \end{array}\right| = \left|\begin{array}{c|c} y & x=f^{-1}(y) \\ \hline 0 & 1 \\ 1 & 0 \\ 1 & 3 \\ 2 & 2 \\ 3 & 4 \\ \dots & \dots \end{array}\right| .$$

It may, or may not, become clear that the new function isn't a function! To make sure, it's a good idea, at the end, to arrange the inputs in the increasing order. Then we clearly see the conflict: $f^{-1}(1) = 0$ and $f^{-1}(1) = 3$. The original function, f, wasn't one-to-one!

The general rule for *finding the inverse* of a function given by a formula follows from the definition:

- ▶ The inverse of $y = f(x)$ is found by solving this equation for x; i.e., $x = f^{-1}(y)$.

This method results in a success only when there is exactly one solution, x, for each y.

Example 3.6.13: inverse of linear polynomial

To find the inverse of a linear polynomial

$$f(x) = 3x - 7\,,$$

set and solve the equation (relation), as follows:

$$y = 3x - 7 \implies y + 7 = 3x \implies \frac{y+7}{3} = x\,.$$

Therefore, we have:

$$f^{-1}(y) = \frac{y+7}{3}\,.$$

If it is not known ahead of time whether the function is one-to-one, this fact is established, automatically, as a part of finding the inverse. For example, to find the inverse of the quadratic function

$$f(x) = x^2\,,$$

we set and solve:

$$y = x^2 \implies \pm\sqrt{y} = x\,.$$

The $\pm$ sign indicates that there are two solutions ($x > 0$). The original function wasn't invertible!

A linear function,

$$f(x) = mx + b\,,$$

is one-to-one and onto whenever $m \neq 0$. Algebraically, we just solve the equation $y = mx + b$ for x. The algebraic result is below.

Theorem 3.6.14: Inverse of Linear Polynomial

The inverse of a linear polynomial

$$f(x) = mx + b,\ m \neq 0\,,$$

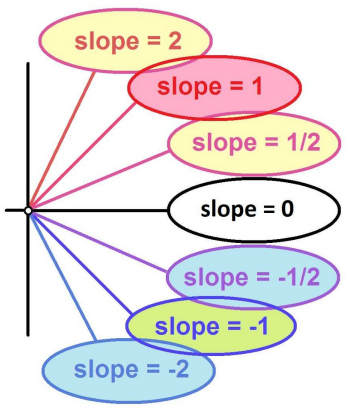
is also a linear polynomial, and its slope is the reciprocal of that of the original:

$$f^{-1}(y) = \frac{1}{m}y - \frac{b}{m}\,.$$

We know that the set of invertible functions is split into pairs of inverses. We can be more specific with the set of all linear polynomials. The pairs have reciprocal slopes, for example:

- 2 and 1/2
- -2 and $-1/2$
- 1 and 1
- -1 and -1
- etc.

We can see these pairs of a steeper line and a shallower line:



Exercise 3.6.15

What do need to do to this sheet of paper in order to make the former land on the latter?

Next, let's try to imagine how some *new algebraic operations* may have emerged.

Some emerged as the *abbreviations* for repeated familiar operations; for example, repeated addition, $2+2+2=2\cdot 3$, leads to a new operation: *multiplication*. Meanwhile, repeated multiplication, $2\cdot 2\cdot 2\cdot 2=2^4$, leads to a new operation: *exponent*. But what about subtraction and division?

Example 3.6.16: subtraction as inverse

Suppose I know how to add. Problem: With \$5 in my pocket, how much do I add to have \$12? Answer: \$7. How do I know? Solve the equation:

$$5+x=12\,.$$

This equation leads to a new operation, *subtraction*: $x=12-5$. Of course, there is also a new function. We can say that "subtraction is the inverse of addition", or more precisely, subtracting 5 is the inverse of adding 5.

Example 3.6.17: division as inverse

Suppose now I know how to multiply. Problem: If I want to make a table 20 inches wide, how many 2-by-4's do I need? Answer: 5. How do I know? Solve the equation:

$$4x=20\,.$$

This equation leads to a new operation, *division*: $x=\dfrac{20}{4}$. Division (by 4) is the inverse of multiplication (by 4).

Example 3.6.18: square root as inverse

Problem: If I want to make a square table with an area 25 square feet, what should be the width of the table? Solve the equation:

$$x\cdot x=25 \implies x^2=25 \implies x=\sqrt{25}\,.$$

Thus, we have a new operation: *square root*. It is the inverse of the squaring function.

Example 3.6.19: cubic root as inverse

Problem: What is the side of a box if its volume is known to be 8 cubic feet? Solve the equation:

$$x^3=8 \implies x=\sqrt[3]{8}\,.$$

The cubic root is the inverse of the cubic power.

Thus, solving equations requires us to *undo* some function present in the equation:

$$\begin{array}{lrcll}
1. & x+2=5 & \implies & (x+2)-2=5-2 & \implies x=3\\
2. & x\cdot 3=6 & \implies & (x\cdot 3)/3=6/3 & \implies x=2\\
3. & x^2=4 & \implies & \sqrt{x^2}=\sqrt{4} & \implies x=2\ (x,y\geq 0)
\end{array}$$

We have "cancellation" on the left and simplification on the right.

We are dealing with functions! And some functions *undo the effect of others*:

1. The addition of 2 is undone by the subtraction of 2, and vice versa.
2. The multiplication by 3 is undone by the division by 3, and vice versa.
3. The second power is undone by the square root (for $x\geq 0$), and vice versa.

Each of these undoes the effect of its counterpart *under substitution*:

1. Substituting $y = x + 2$ into $x = y - 2$ gives us $x = x$.
2. Substituting $y = 3x$ into $x = \frac{1}{3}y$ gives us $x = x$.
3. Substituting $y = x^2$ into $x = \sqrt{y}$ gives us $x = x$, for $x, y \geq 0$.

And vice versa:

1. Substituting $x = y - 2$ into $y = x + 2$ gives us $y = y$.
2. Substituting $x = \frac{1}{3}y$ into $y = 3x$ gives us $y = y$.
3. Substituting $x = \sqrt{y}$ into $y = x^2$ gives us $y = y$, for $x, y \geq 0$.

Warning!

Both cancellations matter.

As we know, it is more precise to say that they undo each other *under composition*: Two numerical functions $y = f(x)$ and $x = g(y)$ are inverse of each other when for every x in the domain of f and for every y in the domain of g, we have:

$$g(f(x)) = x \quad \text{and} \quad f(g(y)) = y\,.$$

This is an alternative way of writing these:

Inverses in substitution notation

$$g(y)\Big|_{y=f(x)} = x \quad \text{and} \quad f(x)\Big|_{x=g(y)} = y$$

Thus, we have three pairs of inverse functions:

$$\begin{aligned} f(x) = x + 2 \quad &\text{vs.} \quad f^{-1}(y) = y - 2 \\ f(x) = 3x \quad &\text{vs.} \quad f^{-1}(y) = \frac{1}{3}y \\ f(x) = x^2 \quad &\text{vs.} \quad f^{-1}(y) = \sqrt{y} \quad \text{for } x, y \geq 0 \end{aligned}$$

Next, it is reasonable to ask: What is the relation between the *graph* of a function and the graph of its inverse?

In other words, what do we need to do with the graph of f to get the graph f^{-1}? The answer is: Hardly anything. After all, a function and its inverse represent the same relation.

The graph of f illustrates how y depends on x – as well as how x depends on y. And the latter is what determines f^{-1}! So, there is no need for a new graph; the graph of f^{-1} *is* the graph of f. The only issue is that the x- and the y-axis point in the wrong directions. It's an easy fix.

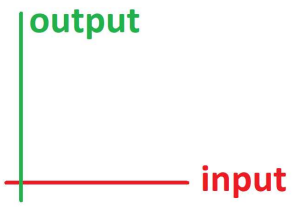
Example 3.6.20: points on the graph of inverse

Suppose we are transitioning from f to its inverse f^{-1}:

$$f = \begin{array}{|c|c|} \hline x & y = f(x) \\ \hline 2 & 5 \\ 3 & 1 \\ 8 & 7 \\ \cdots & \\ \hline \end{array} \quad \Longrightarrow \quad f^{-1} = \begin{array}{|c|c|} \hline y & x = f^{-1}(y) \\ \hline 5 & 2 \\ 1 & 3 \\ 7 & 8 \\ \cdots & \\ \hline \end{array}$$

These are the same pairs! Therefore, they are represented by the same points on the xy-plane.

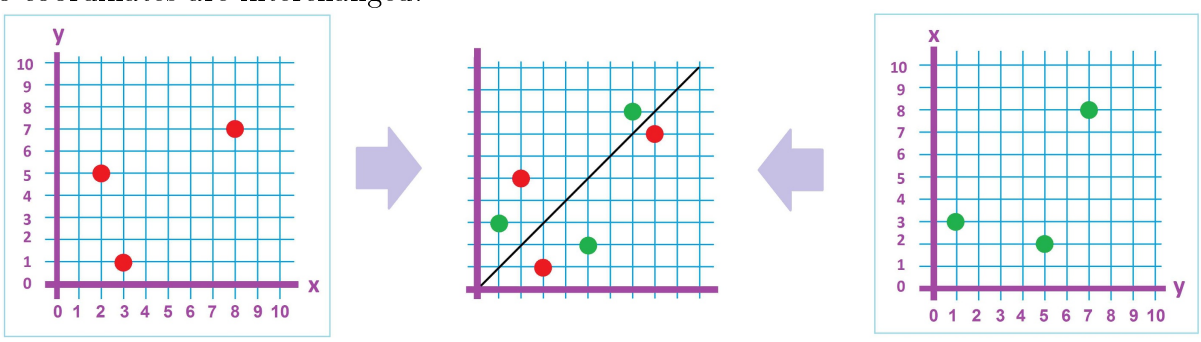
It is common, however, to put the input variable in the horizontal axis and the output in the vertical.



This makes us replace the points in the xy-plane with new points in the yx-plane:

(x, y)	$\longrightarrow$	(y, x)
$(2, 5)$	$\longrightarrow$	$(5, 2)$
$(3, 1)$	$\longrightarrow$	$(1, 3)$
$(8, 7)$	$\longrightarrow$	$(7, 8)$
...		...

The two coordinates are interchanged:



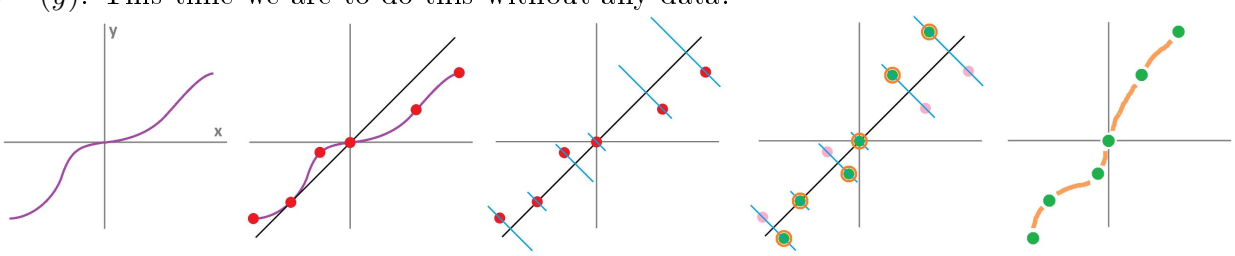
We realize that each point jumps across the diagonal line $y = x$! So, we have a match:

▶ Every point (x, y) in the xy-plane corresponds to the point (y, x) in the yx-plane.

Above we made a copy of the graph of f, flipped it, and then on top of the original.

Example 3.6.21: inverse graph point by point

Suppose, again, a function is given only by its graph and we need to construct the graph of the inverse $x = f^{-1}(y)$. This time we are to do this without any data:



Start with choosing a few points on the graph. Each of them will jump across the diagonal under this flip. How exactly? The general rule for plotting a counterpart of a point is the following:

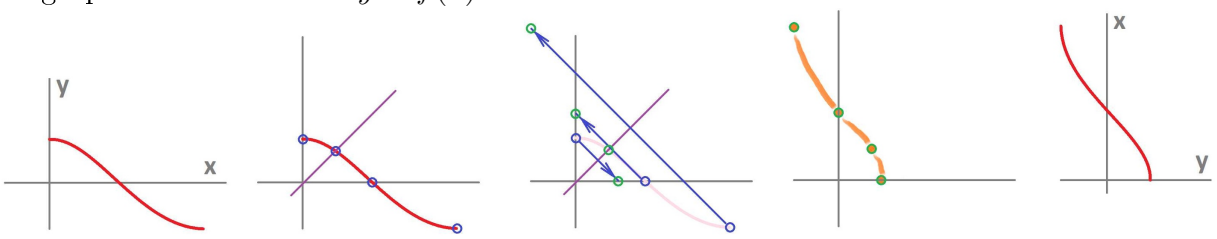
▶ From the point go perpendicular to the diagonal and then measure the same distance on the other side.

In other words, we plot a line through our point with slope -1.

Now, we can simplify our job by choosing the points more judiciously; we choose ones with easy-to-find counterparts. First, points on the diagonal don't move by the flip about the diagonal. Second, points on one of the axes jump to the other axis with no need for measuring. Finally, once all points are in place, finally, draw a curve that connects them.

Example 3.6.22: graph of inverse point by point

Plot the graph of the inverse of $y = f(x)$ shown below:

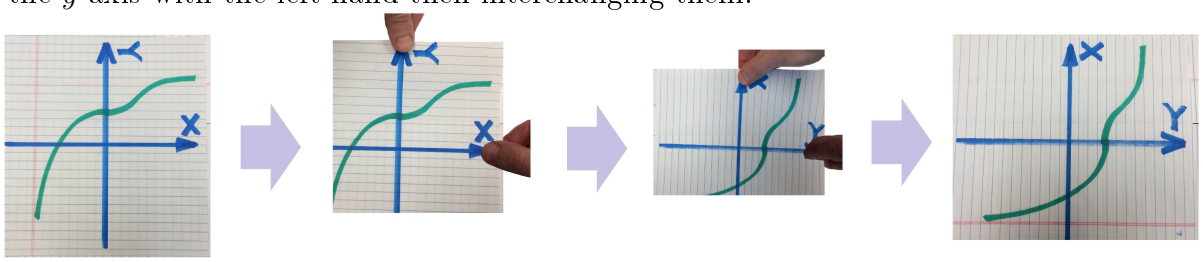


These are the steps:

- Draw the diagonal $y = x$.
- Pick a few points on the graph of f (we choose four).
- Plot a corresponding point for each of them:
  - on the line through point A that is perpendicular to the diagonal (i.e., its slope is 45 degrees down)
  - on the other side of the diagonal from A
  - at the same distance from the diagonal as A
- Draw by hand a curve from point to point.

Example 3.6.23: flip graph

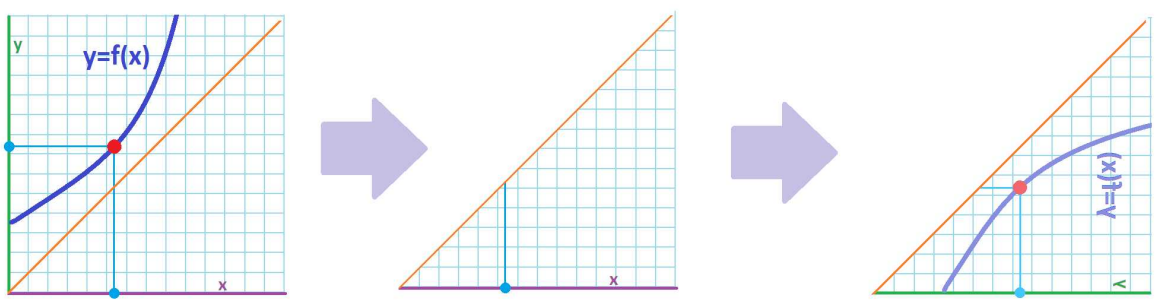
The following can be done *without a pen*. If we have a piece of paper with the xy-axis and the graph of $y = f(x)$ on it, we flip it by grabbing the end of the x-axis with the right hand and grabbing the end of the y-axis with the left hand then interchanging them:



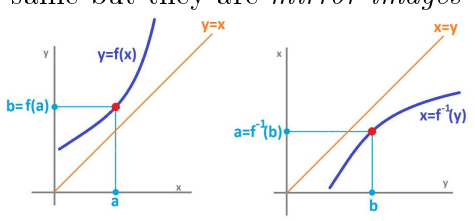
We face the *opposite side* of the sheet then, but the graph is still visible: the x-axis is now pointing up and the y-axis right, as intended. A transparent sheet of plastic would work even better.

Example 3.6.24: fold graph

Alternatively, we can also fold:



The shapes of the graphs are the same but they are *mirror images* of each other:



Exercise 3.6.25

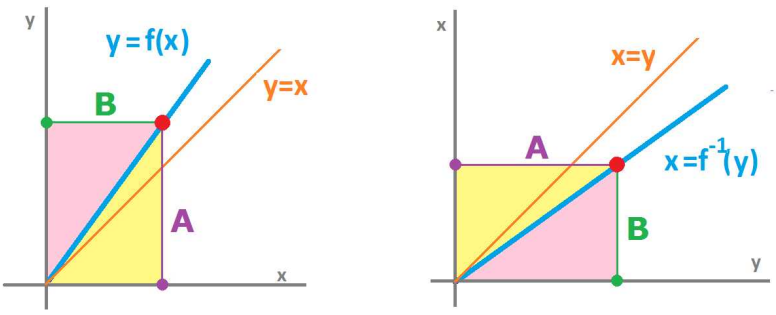
What graphs will land on themselves under this transformation?

Exercise 3.6.26

What letters have this kind of symmetry?

Example 3.6.27: slope of inverse

This is what happens when we apply this flip to the graph of a linear function:



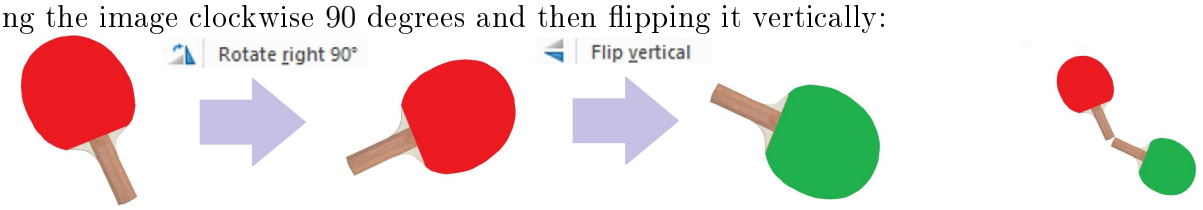
Then, we can compare:

$$\text{slope of } f \;=\; \frac{\text{rise}}{\text{run}} = \frac{A}{B} \quad \text{and} \quad \text{slope of } f^{-1} \;=\; \frac{\text{rise}}{\text{run}} = \frac{B}{A}\,.$$

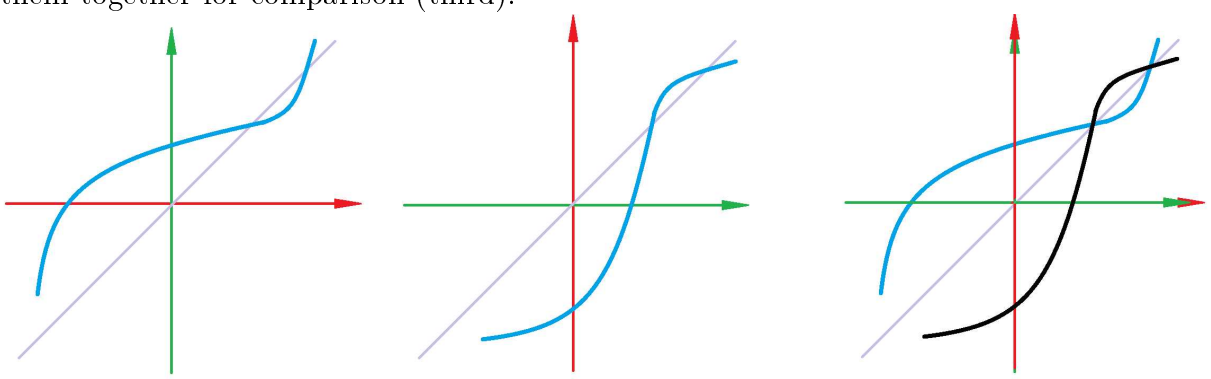
They are, as we already know, the reciprocals of each other!

Example 3.6.28: inverse graph with computer graphics

Such a transformation of the plane can be accomplished with simple image editing software by first rotating the image clockwise 90 degrees and then flipping it vertically:



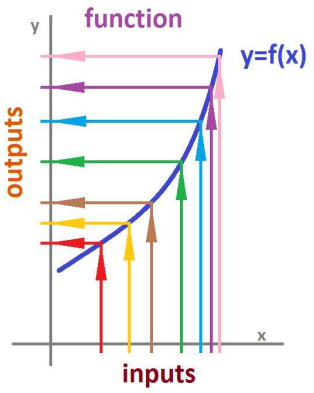
This is how starting from a graph (first below), we find the graph of the inverse (second), and then bring them together for comparison (third):

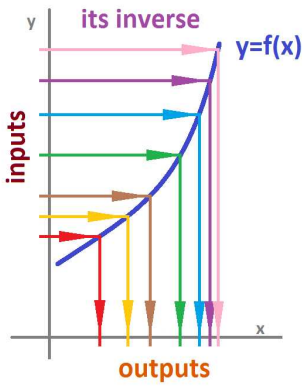


Warning!

It's ill-advised to try to guess what the graph of the inverse looks like.

Remember, we only need the graph of the original function to be able to evaluate all the values of the inverse:





Next, some more profound issues...

Example 3.6.29: square vs. square root

Here is a familiar pair of functions that undo each other:

$$x \;\rightarrow\; \boxed{\text{square}} \;\rightarrow\; y \;\rightarrow\; \boxed{\text{square root}} \;\rightarrow\; x\text{, same?}$$

Now, the diagram fails if we plug in $x = -2$:

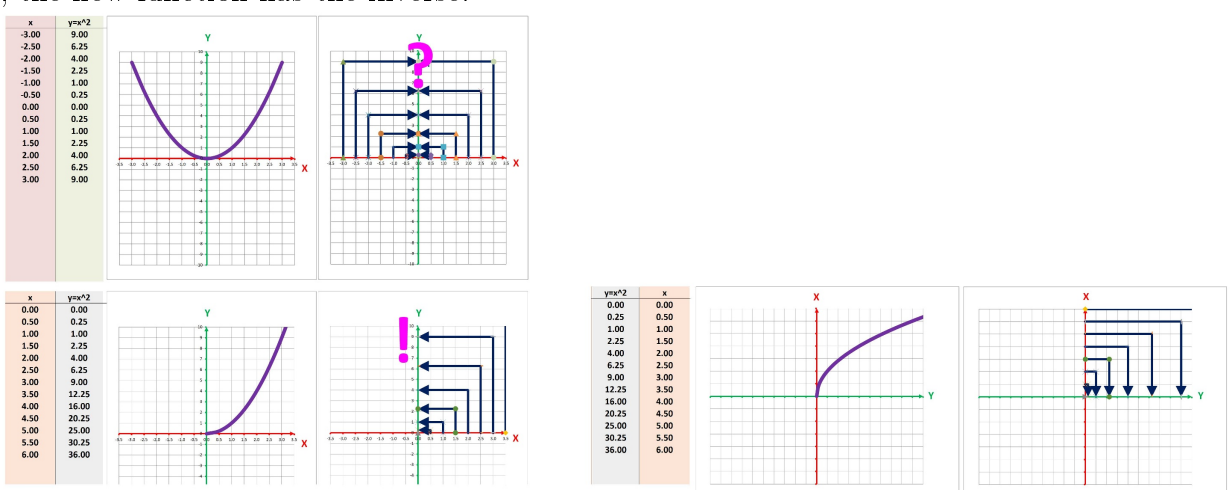
$$-2 \;\rightarrow\; \boxed{\text{square}} \;\rightarrow\; 4 \;\rightarrow\; \boxed{\text{square root}} \;\rightarrow\; 2\text{, not the same!}$$

As we know, not all functions have inverses...

However, we can make it work by *restricting the domain and the codomain* of the function:

- The old function is $y = x^2$, with the domain and the codomain assumed to be $(-\infty, \infty)$ (no inverse).
- The new function is $y = x^2$, with the domain and the codomain chosen to be $[0, \infty)$.

Then, the new function has the inverse:



We have made it possible by removing the second possibility:

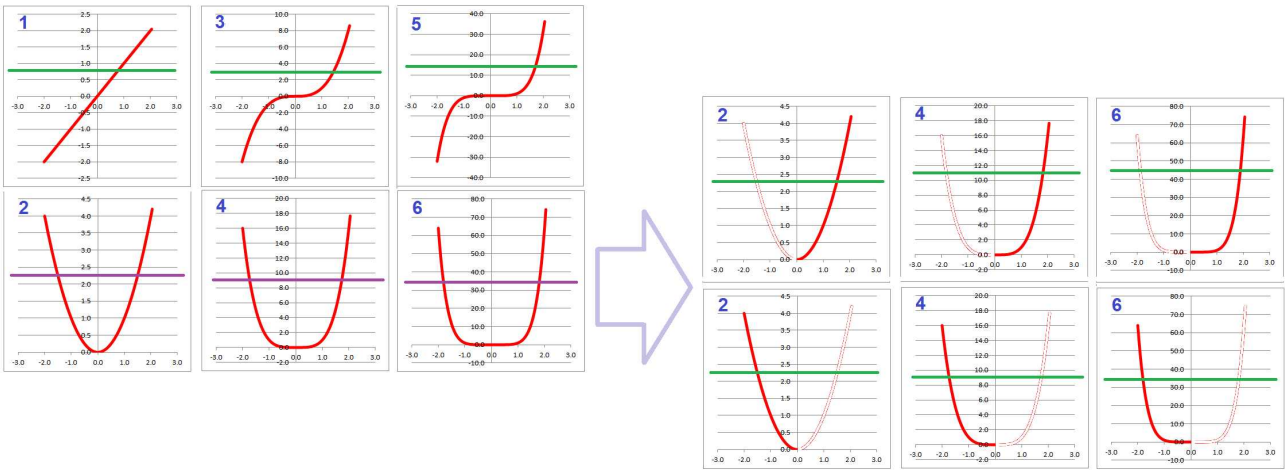
$$2^2 = 4,\ (-2)^2 = 4\,.$$

The following will be very useful.

Theorem 3.6.30: Classification of Power Functions

1. *The odd powers are one-to-one; consequently, they are invertible. The even powers aren't one-to-one; consequently, they are not invertible.*
2. *The even powers with domains and codomains reduced to $[0, +\infty)$ (or $(-\infty, 0]$) are one-to-one; consequently, they are invertible.*

We can see below how the *Horizontal Line Test* is satisfied for the odd powers, fails for the even powers, and is satisfied for the "reduced" even powers:



Warning!

We will have to keep track of both branches (separately) when we solve equations:

$$x^2 = 1 \implies x = 1 \quad \texttt{OR} \quad x = -1\,.$$

Example 3.6.31: inverse from formula

Suppose this time that a function is given only by its *formula*. Find the formula of the inverse $x = f^{-1}(y)$. For example, let

$$f(x) = x^3 - 3\,.$$

We simply rewrite

$$y = x^3 - 3\,,$$

and then solve for x:

$$x = \sqrt[3]{y+3}\,.$$

Thus, the answer is:

$$f^{-1}(y) = \sqrt[3]{y+3}\,.$$

More on solving equations in Chapter 5.

Exercise 3.6.32

Function $y = f(x)$ is given below by a list its values. Find its inverse and represent it by a similar table.

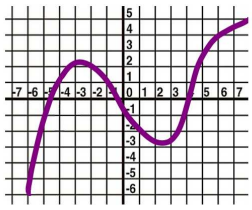
x	0	1	2	3	4
$y = f(x)$	1	2	0	4	3

Exercise 3.6.33

What kind of function is its own inverse?

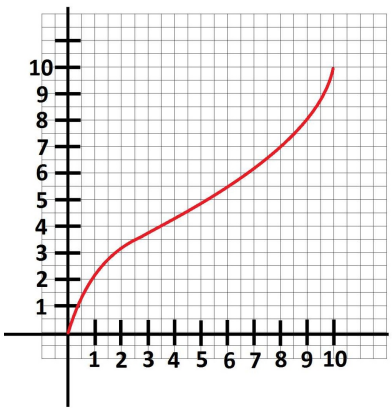
Exercise 3.6.34

Plot the inverse of the function shown below, if possible:



Exercise 3.6.35

Plot the graph of the inverse of this function:



Exercise 3.6.36

Function $y = f(x)$ is given below by a list of its values. Is the function one-to one? What about its inverse?

x	0	1	2	3	4
$y = f(x)$	7	5	3	4	6

Exercise 3.6.37

Plot the graph of the function $f(x) = \dfrac{1}{x-1}$ and the graph of its inverse. Identify its important features.

3.7. Units conversions and changes of variables

The variables of the functions we are considering are quantities we meet in everyday life. Frequently, there are multiple ways to measure these quantities:

- length and distance: inches, miles, meters, kilometers, ..., light years
- area: square inches, square miles, ..., acres
- volume: cubic inches, cubic miles, ..., liters, gallons
- time: minutes, seconds, hours, ..., years
- weight: pounds, grams, kilograms, karats
- temperature: degrees of Celsius, of Fahrenheit
- money: dollars, euros, pounds, yen
- etc.

Almost all conversion formulas are just multiplications, such as this one:

$$\#\text{ of meters} = \#\text{ of kilometers} \cdot 1000\,.$$

Warning!

We don't convert "pounds to kilos", we convert the *number of* pounds to the *number of* kilos.

The only exception of the temperature, because 0 degrees of Celsius doesn't correspond to 0 degrees of Fahrenheit.

This is the relation between degrees and radians:

$$\pi \text{ radians} = 180 \text{ degrees}\,.$$

In other words, the conversion between the number of degrees d and the number of radians r is the following relation:

$$\pi r = 180 d\,.$$

Therefore, we convert from degrees to radians with the following function:

$$r = \frac{\pi}{180} d\,.$$

Then, we convert from radians degrees with the inverse of this function:

$$d = \frac{180}{\pi} r\,.$$

Within each of the categories, there may be complex, even circular, relations. For example, we have the following among these currencies:

$$\begin{array}{ccc} \#\text{ of dollars} & \xrightarrow{\times .9} & \#\text{ of euros} \\ \uparrow{\scriptstyle \times 1.3} & & \downarrow{\scriptstyle \times 122} \\ \#\text{ of pounds} & \xleftarrow{\times 0.007} & \#\text{ of yen} \end{array} \qquad \text{or} \qquad \begin{array}{ccc} USD & \xleftarrow{/.9} & EUR \\ \downarrow{\scriptstyle /1.3} & & \uparrow{\scriptstyle /122} \\ GBP & \xrightarrow{/0.007} & JPY \end{array}$$

The arrows, of course, indicate functions, two in a row indicate compositions, and the reversed arrows are the inverses!

Exercise 3.7.1

Make your way from minutes to years.

We don't deal with these quantities one by one nor even in these pairs. We will study the *functions* that have them as variables.

We will first consider the *compositions* of these functions with the functions that represent the unit conversions.

Example 3.7.2: units of distance

Suppose t is the time and x is the location. Suppose also that a function g represents the change of units of length, such as from miles to kilometers:

$$z = g(x) = 1.6x\,.$$

Then, the change of the units will make very little difference; the coefficient, $m = 1.6$, is the only adjustment necessary. If f is the distance in miles, then h is the distance in kilometers: $h(t) = 1.6f(t)$. Thus, all the functions are replaced with their multiples. The graphs are stretched!

We call such a unit conversion a *change of variables*. Usually, it is done one at a time: either the dependent or the independent variable.

Example 3.7.3: motion and units

Suppose we study *motion* and we have a function $y = f(x)$ that relates

- x, time in minutes, to
- y, location in inches.

What if we need to switch to

- t, time in seconds, or
- z, location in feet?

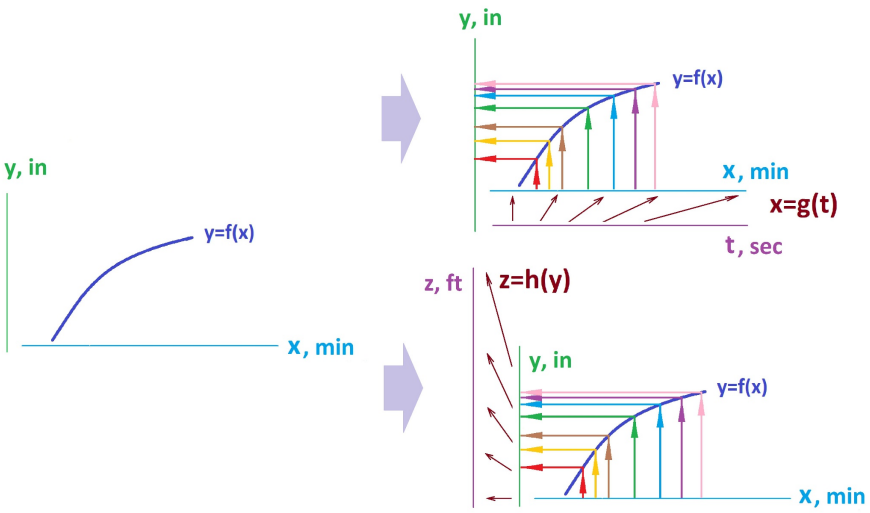
The algebra is clear:

$$x = t/60 \quad \text{and} \quad z = y/12\,.$$

Then we might have two new functions:

$$y = f(t/60) \quad \text{and} \quad z = f(x)/12\,.$$

Now, what will the new graphs look like? To answer, we combine the graph of f with the two transformations of the two axes, as follows:



The result is a vertical and a horizontal stretch/shrink. However, it's entirely up to us to choose the units on the new axes to match the old: the graph will remain the same!

Exercise 3.7.4

What is the relation between seconds and feet?

Example 3.7.5: time and temperature

Suppose we have a function f that records the temperature (in Fahrenheit) as a function of time (in minutes).



Question:

▶ What should f be replaced with if we want to record the temperature in Celsius as a function of time in seconds?

Let's name the variables:

- s is the time in seconds,
- m is the time in minutes,
- F is the temperature in Fahrenheit,
- C is the temperature in Celsius.

Suppose the original function, say,

$$F = f(m)\,,$$

is to be replaced with some new function,

$$C = g(s)\,.$$

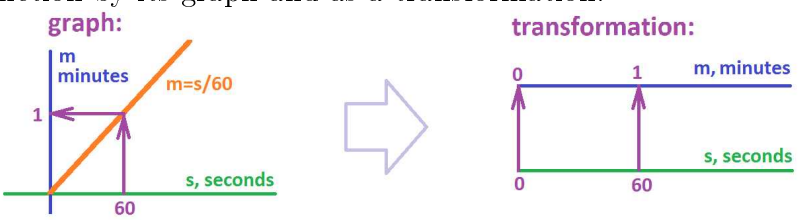
First, we need the *conversion formulas* for these units. First, the time. This is what we know:

$$1 \text{ minute } = 60 \text{ seconds.}$$

However, this is not the formula to be used to convert s to m because these are the *number* of seconds and the *number* of minutes, respectively. Instead, we have

$$m = s/60\,.$$

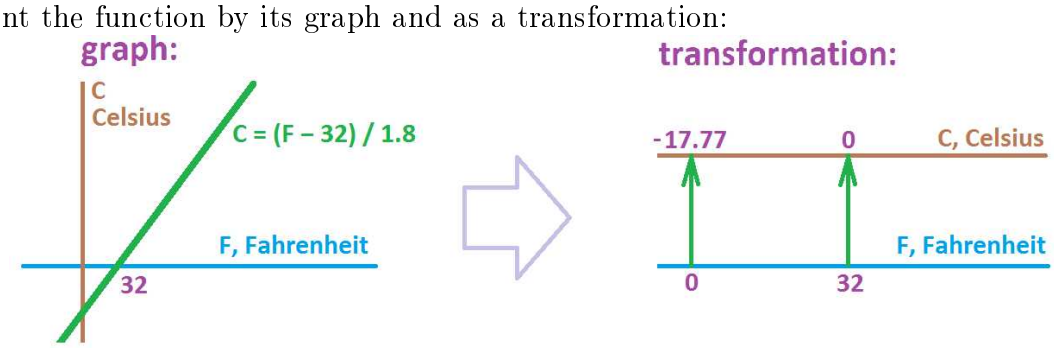
We represent the function by its graph and as a transformation:



Second, the temperature. This is what we know:

$$C = (F - 32)/1.8\,.$$

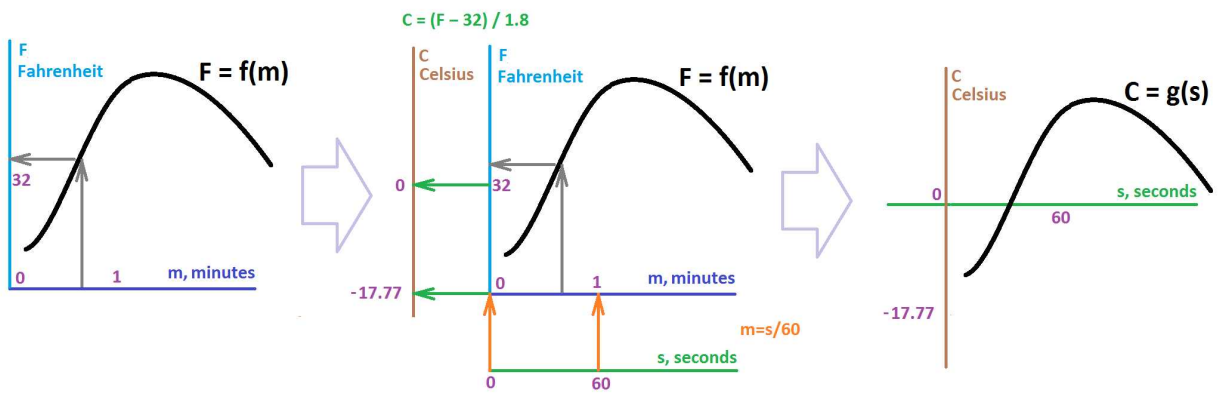
We represent the function by its graph and as a transformation:



These are the relations between the four quantities:

$$g: \quad s \xrightarrow{\;s/60\;} m \xrightarrow{\;f\;} F \xrightarrow{\;(F-32)/1.8\;} C$$

Instead of transforming the axes and, therefore, the plane, we choose to simply *relabel* them:



The answer to our question is, we replace f with g, the *composition* of the above functions:

$$F = g(s) = \big(f(s/60) - 32\big)/1.8\,.$$

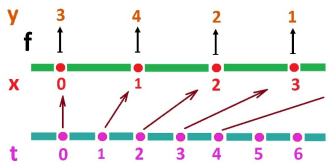
Note that both of the conversion formulas are one-to-one functions! That's what guarantees that the conversions are unambiguous and reversible. More precisely, we say that these functions are *invertible*. Indeed, these are the inverses, for the time:

$$s = 60m\,,$$

and for the temperature:

$$F = 1.8C + 32\,.$$

Note that all of the conversion formulas have been provided by *linear functions*. Then, a linear change of variables will cause the x-axis or the y-axis to shift, stretch, or flip (vertically or horizontally):



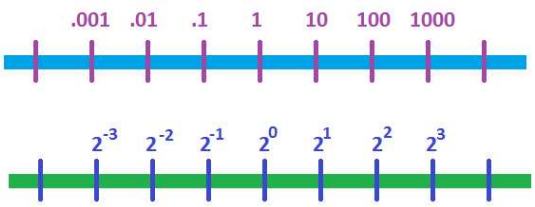
We conclude:

- A linear change of variables will cause the graph of the function to shift, stretch, or flip.

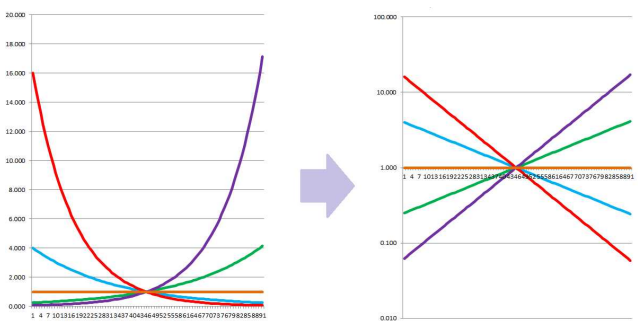
Some nonlinear changes of variables are also known.

Example 3.7.6: non-linear change of units

The loudness of sound (the decibel) and the magnitude of earthquakes (the Richter scale) are measured on a *logarithmic scale*. This scale is based on orders of magnitude rather than a linear scale, i.e., the next mark on the scale corresponds to the magnitude at the previous mark multiplied by a predetermined coefficient, such as 10:



Spreadsheet software may have an option to switch the scale with just a couple of clicks. For example, we can see how the graph of a geometric progression becomes a *straight line* when we switch the y-axis to the logarithmic scale:



Exercise 3.7.7

Establish algebraic relation between the quantities expressed in the units listed in the beginning of the section.

Exercise 3.7.8

What is the relation between miles per gallon and kilometers per liter?

Exercise 3.7.9

What is the relation between miles per gallon and gallons per mile?

Example 3.7.10: what happens to formulas under conversions of units

What happens to the formula of a function under conversions of units and transformations of the axes? We just execute the corresponding substitutions in the formula.

For example, consider a function

$$y = f(x)\,.$$

What will be the formula of this function if we

- shift it 5 units right, and then
- shift it 2 units up?

We have new variables:

$$u = x + 5,\ v = y + 2\,.$$

We need to execute the change of variables via a substitution. For that, we solve for x and y:

$$x = u - 5,\ y = v - 2\,.$$

We now substitute these into the function

$$v - 2 = f(u - 5)\,.$$

We have a new *relation* between the new variables, but for a *function*, we need an explicit formula. We solve for v:

$$v = f(u - 5) + 2\,.$$

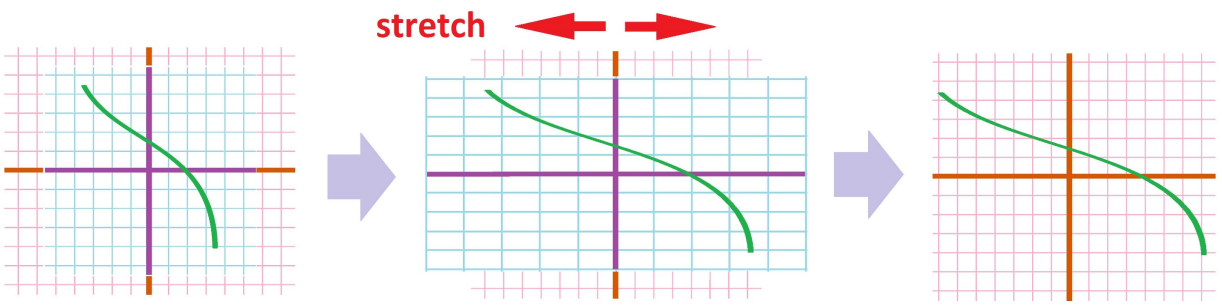
That's the equation of the new curve on the new uv-plane. Compare these, however:

- The shift 2 units *up* has produced $+2$ in the formula, as expected.
- The shift 5 units *right* has produced -5 in the formula, the opposite of what's expected.

This minus sign for the independent variable, unlike the dependent variable, is something to watch out for when dealing with transformations of *formulas* of functions.

3.8. Transforming the axes transforms the plane

In the context of our study of numerical functions, why do we care about transformations of the plane? Because their graphs are drawn on the xy-plane:

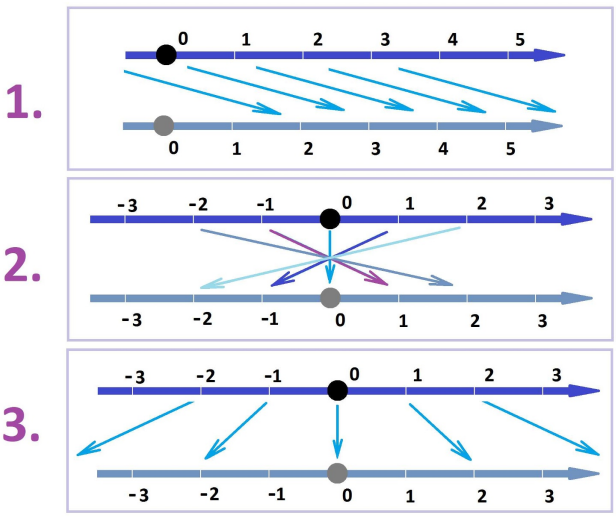


We also know that functions transform the real line and, therefore, transform the *axes* of the xy-plane.

We narrow this down:

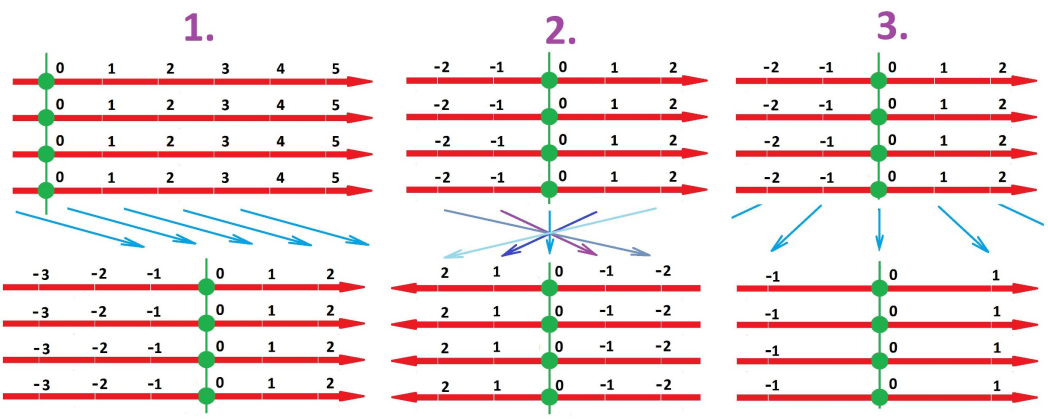
▶ How do the transformations of the axes – horizontal and vertical – affect the xy-plane?

Let's review what we know. These are the three linear transformations of *an* axis: shift, flip, and stretch:



Now, let's imagine that transforming an axis transforms – in unison – all the lines on the plane parallel to it.

The x-axis and its transformations are *horizontal* and so are the transformations of the xy-plane:

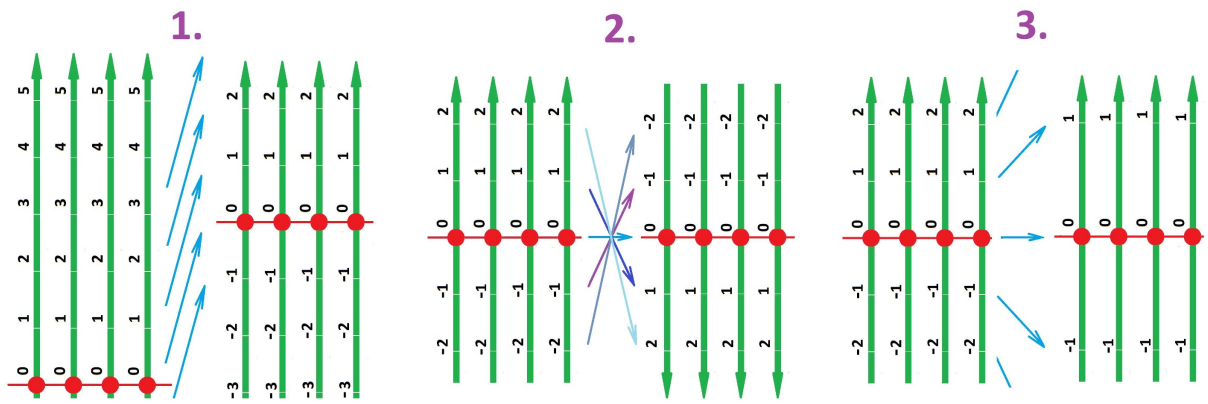


Then, the shift of the x-axis becomes a horizontal shift of the xy-plane, the flip of the x-axis becomes a horizontal flip of the xy-plane (around the vertical axis), and the stretch of the x-axis becomes a horizontal stretch of the xy-plane (away from the vertical axis).

For an algebraic representation of these transformations, we just add y, that remains unchanged, to the formula, as follows:

$$\begin{array}{lllll}
x & \xrightarrow{\text{shift by } s} & x+s & \Longrightarrow (x,y) & \xrightarrow{\text{horizontal shift by } s} (x+s,y) \\
x & \xrightarrow{\text{flip}} & -x & \Longrightarrow (x,y) & \xrightarrow{\text{horizontal flip}} (-x,y) \\
x & \xrightarrow{\text{stretch by } k} & x\cdot k & \Longrightarrow (x,y) & \xrightarrow{\text{horizontal stretch by } k} (x\cdot k,y)
\end{array}$$

What about the y-axis? The y-axis and its transformations are *vertical* and so are the transformations of the xy-plane:



The shift of the y-axis produces a vertical shift of the xy-plane, the flip of the y-axis produces a vertical flip of the xy-plane (around the horizontal axis), and the stretch of the y-axis produces a vertical stretch of the xy-plane (away from the horizontal axis).

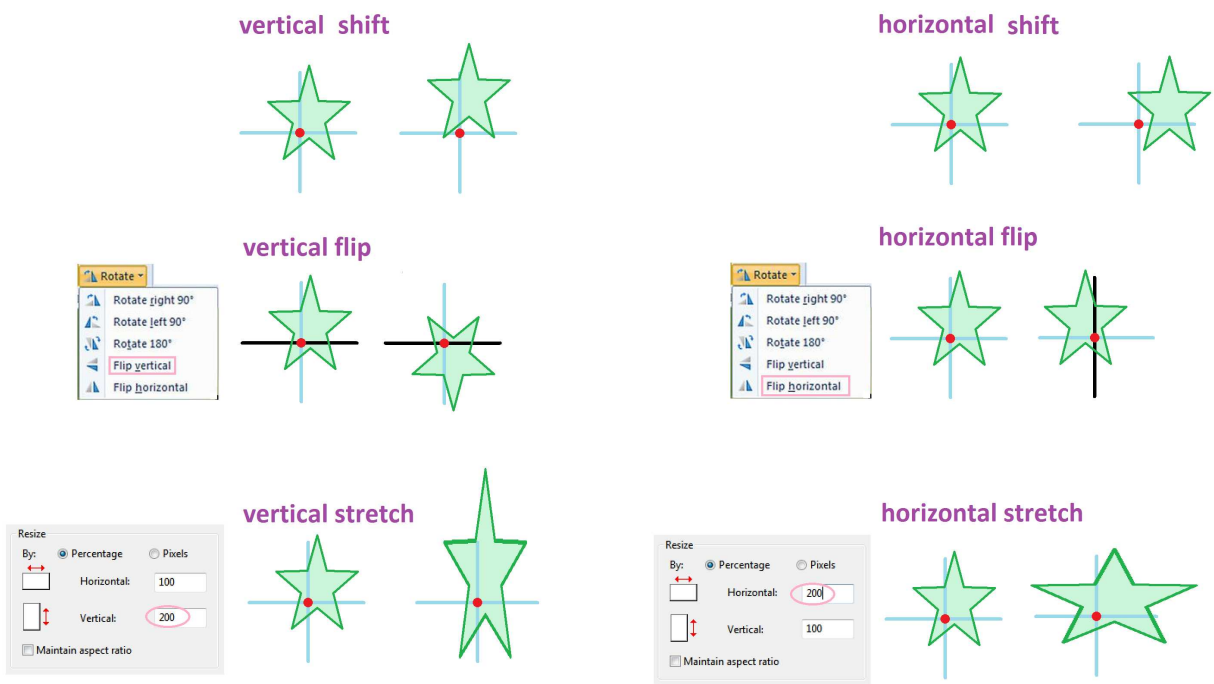
For an algebraic representation of these transformations, we just add x, that remains unchanged, to the formula:

$$\begin{array}{lllll} y & \xrightarrow{\text{shift by } s} & y+s & \Longrightarrow (x,y) & \xrightarrow{\text{vertical shift by } s} (x,y+s) \\ y & \xrightarrow{\text{flip}} & -y & \Longrightarrow (x,y) & \xrightarrow{\text{vertical flip}} (x,-y) \\ y & \xrightarrow{\text{stretch by } k} & y\cdot k & \Longrightarrow (x,y) & \xrightarrow{\text{vertical stretch by } k} (x,y\cdot k) \end{array}$$

Horizontal transformations don't change y and vertical don't change x!

Example 3.8.1: transformations with computer graphics

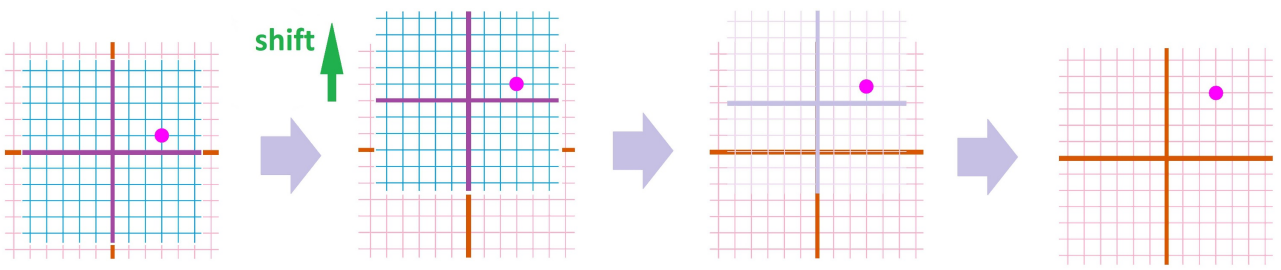
We can illustrate these transformations with a graphics editor:



The first two rows show *rigid motions*, while the last is re-scaling.

So, the algebra of the real line creates a new algebra of the Cartesian plane. Let's revisit these six transformations one by one.

We start with a *vertical shift*. We shift the whole xy-plane as if it is printed on a sheet of paper. Furthermore, there is another sheet of paper underneath used for reference. It is to the second sheet that we transfer the resulting points. We then use its coordinate system to record the coordinates of the new point. For example, a shift of 3 units upward is shown below:



So, all vertical lines are shifted up by s. Then, the whole plane is shifted $s > 0$ units up. A generic point (x, y) makes a step up/down by s and becomes $(x, y + s)$. This is another algebraic way to present the transformation:

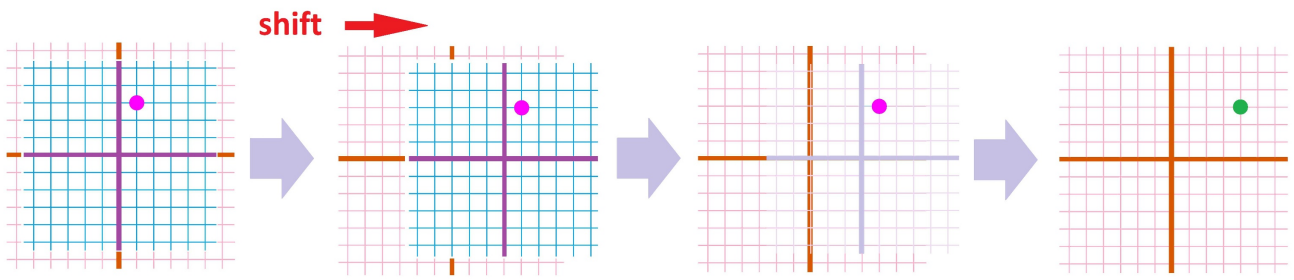
$$(x, y) \xrightarrow{\text{up } s} (x, y + s)$$

It is as if the algebra of the flip of the y-axis given previously, $y \mapsto y + s$, is copied and then paired up with x.

Exercise 3.8.2

What is the effect of two vertical stretches executed consecutively?

What about the *horizontal shift*? For example, a shift of 2 units right is shown below:



So, all horizontal lines are shifted right by s. Then, the whole plane is shifted $s > 0$ units right. A generic point (x, y) makes a step right/left by s and becomes $(x + s, y)$. This is an algebraic way to present the transformation:

$$(x, y) \xrightarrow{\text{right } s} (x + s, y)$$

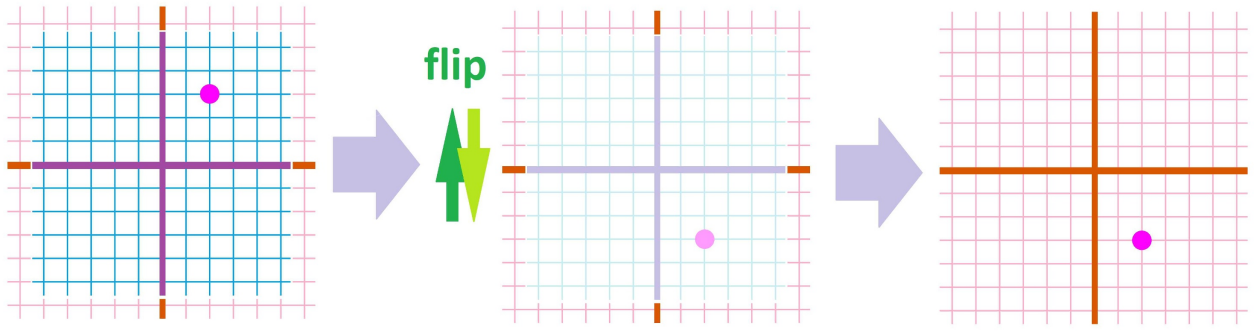
It is as if the algebra of the flip of the x-axis given previously, $x \to x + s$, is copied and then paired up with y.

Exercise 3.8.3

What is the effect of a vertical stretch and a horizontal stretch executed consecutively? What if we change the order?

These shifts can also be described as *a translation along the y-axis* and *a translation along the x-axis*, respectively.

Now a *vertical flip*. We lift, then flip the sheet of paper with the xy-plane on it, and finally place it on top of another such sheet so that the x-axes align. This flip is shown below:



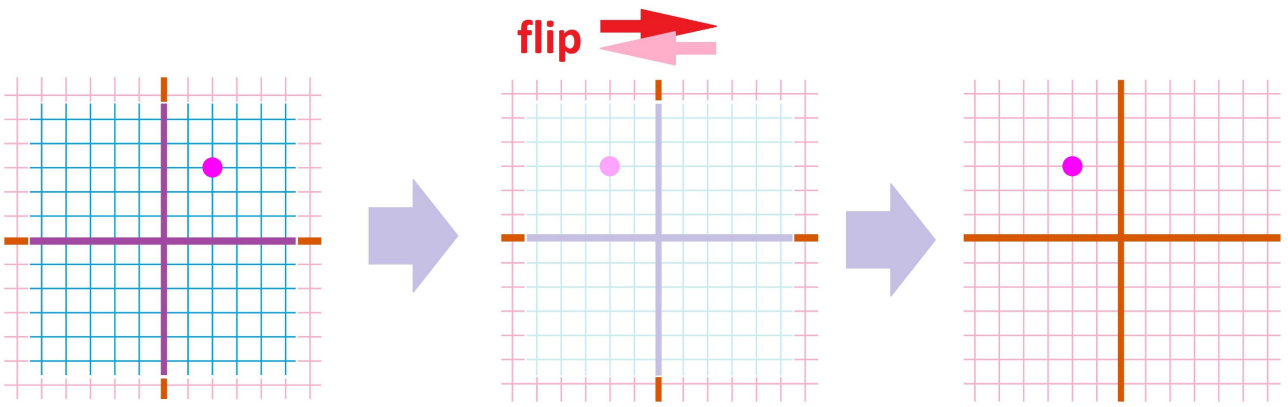
So, all vertical lines are flipped about their origins. Then, the whole plane is flipped about the x-axis. A generic point (x, y) jumps across the x-axis and becomes $(x, -y)$. This is the algebraic outcome:

$$(x, y) \xrightarrow{\text{vertical flip}} (x, -y)$$

Exercise 3.8.4

What is the effect of two vertical flips executed consecutively?

For the *horizontal flip*, we lift, then flip the sheet of paper with the xy-plane on it, and finally place it on top of another such sheet so that the y-axes align. This flip is shown below:



So, all horizontal lines are flipped about their origins. Then, the whole plane is flipped about the y-axis. A generic point (x, y) jumps across the y-axis and becomes $(-x, y)$. This is the algebraic outcome:

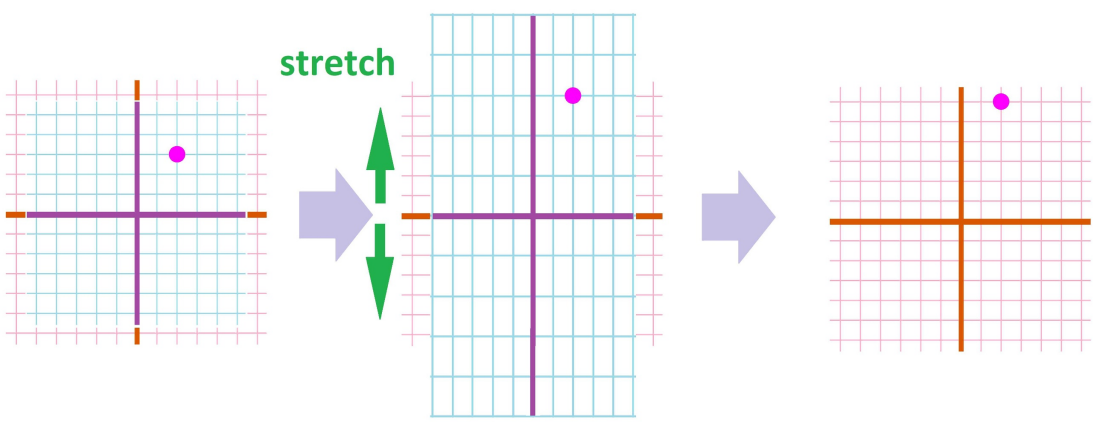
$$(x, y) \xrightarrow{\text{horizontal flip}} (-x, y)$$

Exercise 3.8.5

What is the effect of a vertical flip and a horizontal flip executed consecutively? What if we change the order?

These flips can also be described as *a mirror reflection about the x-axis* and *a mirror reflection about the y-axis*, respectively.

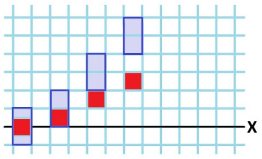
Next, a *vertical stretch*. The coordinate system isn't on a piece of paper anymore! It is on a rubber sheet. We grab it by the top and the bottom and pull them apart in such a way that the x-axis doesn't move. For example, a stretch by a factor of 2 is shown below:



So, all vertical lines are stretched by $k > 0$ away from their origins. Then, the whole plane is stretched by a factor k away from the x-axis. The distance of a generic point (x, y) from the x-axis grows proportionally to k and the point becomes $(x, y \cdot k)$. This is the algebra to describe it:

$$(x, y) \xrightarrow{\text{vertical stretch by } k} (x, y \cdot k)$$

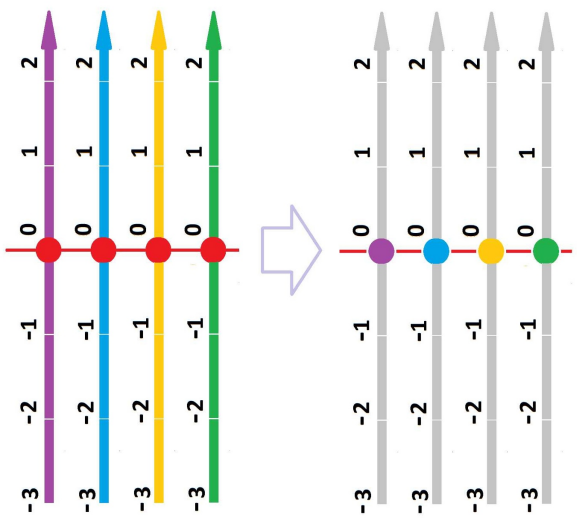
Even though the stretch is the same for all subsets of the plane, the new location will vary depending on the location of the subset relative to the x-axis:



Exercise 3.8.6

What is the effect of a vertical flip and a horizontal shift executed consecutively? What if we change the order?

The case $k = 0$ is very special. As each vertical line collapses on its x-intercept, the whole plane lands on the x-axis. It is called the *projection on the x-axis*:



This is the algebra:

$$(x, y) \xrightarrow{\text{projection}} (x, 0)$$

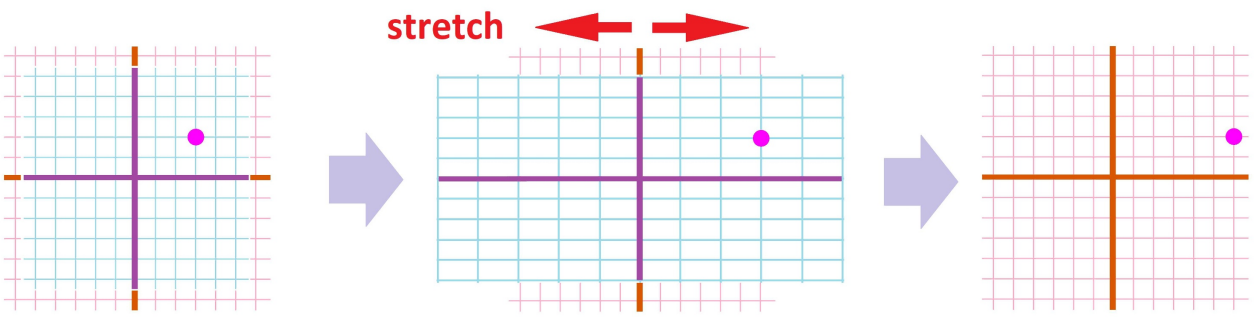
Exercise 3.8.7

What is the range of the projection?

Exercise 3.8.8

What is the effect of a vertical flip and a projection on the x-axis executed consecutively? What if we change the order?

What about *horizontal stretch*? This time, we grab it by the left and right edges of the rubber sheet and pull them apart in such a way that the y-axis doesn't move. For example, a stretch by a factor of 2 is shown below:



So, all horizontal lines are stretched by $k > 0$ away from their origins. Then, the whole plane is stretched by a factor k away from the y-axis. The distance of a generic point (x, y) from the y-axis grows proportionally

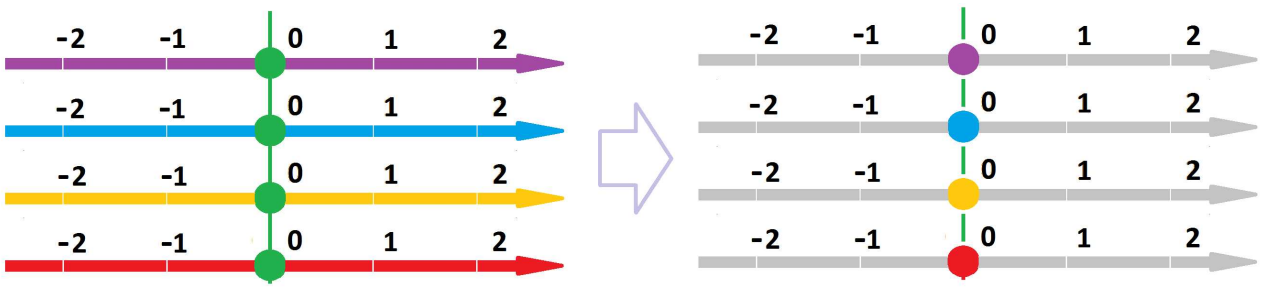
to k and the point becomes (kx, y). This is a way describe a horizontal stretch:

$$(x, y) \xrightarrow{\text{horizontal stretch by } k} (x \cdot k, y)$$

Exercise 3.8.9

What is the effect of a vertical flip and a horizontal flip executed consecutively? What if we change the order?

The case $k = 0$ is very special. As each horizontal line collapses on its y-intercept, the whole plane lands on the y-axis. It is called the *projection on the y-axis*:



This is the algebra:

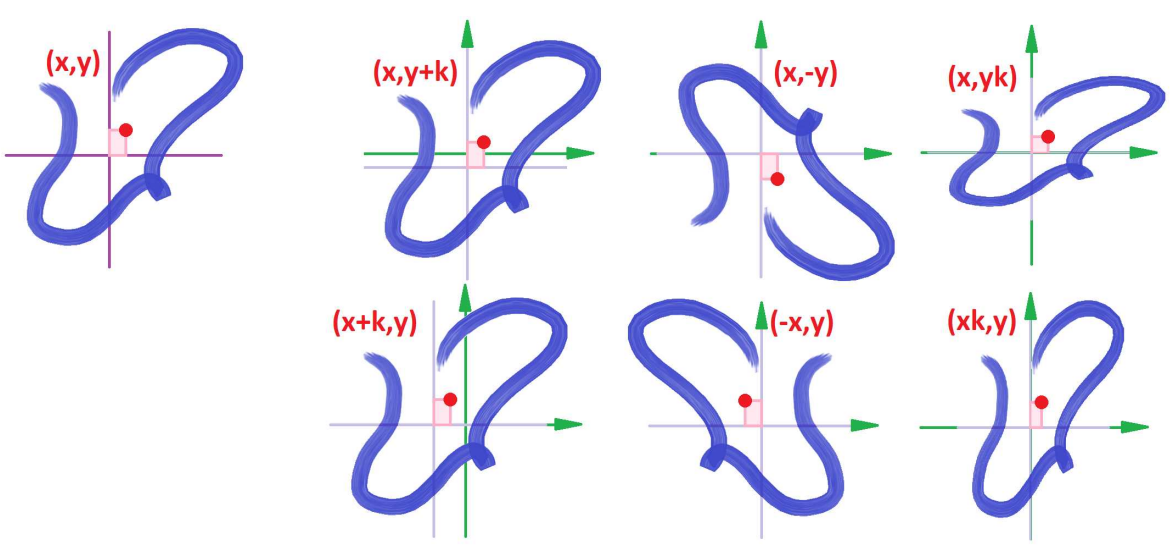
$$(x, y) \xrightarrow{\text{projection}} (0, y)$$

Exercise 3.8.10

What is the effect of a vertical projection and a horizontal projection executed consecutively?

These stretches can also be described as *a uniform deformation away from the y-axis* and *a uniform deformation away from the y-axis*, respectively.

These are our six basic transformations:



The algebra below reflects the geometry above.

Theorem 3.8.11: Formulas of Transformations of Plane

The following transformations of the plane are given by their formulas:

vertical shift:	*flip:*	*stretch:*
$\begin{pmatrix} x & , & y \end{pmatrix}$	$\begin{pmatrix} x & , & y \end{pmatrix}$	$\begin{pmatrix} x & , & y \end{pmatrix}$
$\mapsto \begin{pmatrix} x & , & y+k \end{pmatrix}$	$\mapsto \begin{pmatrix} x & , & y \cdot (-1) \end{pmatrix}$	$\mapsto \begin{pmatrix} x & , & y \cdot k \end{pmatrix}$

horizontal shift:	*flip:*	*stretch:*
$(\ x\ ,\ y\ )$	$(\ x\ ,\ y\ )$	$(\ x\ ,\ y\ )$
$\mapsto (\ x+k\ ,\ y\ )$	$\mapsto (\ x\cdot(-1)\ ,\ y\ )$	$\mapsto (\ x\cdot k\ ,\ y\ )$

It will be important later that these transformations of the plane are *functions of the plane to itself*, i.e.,

$$F : \mathbf{R}^2 \to \mathbf{R}^2 .$$

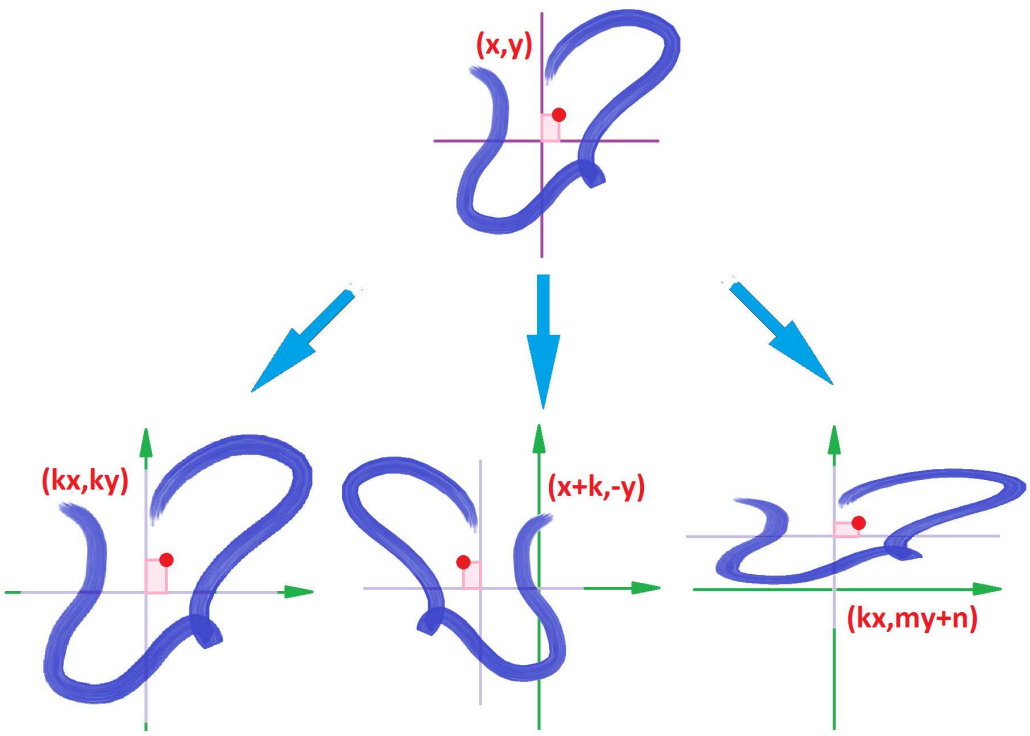
Exercise 3.8.12

What are the images of these six functions? What about the projections?

For now, each of these six operations is limited to one of the two directions: along the x-axis or along the y-axis. We *combine* them as compositions. For example,

point	$\to$	stretch vertically by k	$\to$	point	$\to$	flip horizontally	$\to$	point
(x,y)	$\to$	multiply y by k	$\to$	(x,yk)	$\to$	multiply x by (-1)	$\to$	$(-x,yk)$

We produce a variety of results:



Exercise 3.8.13

Execute – both geometrically and algebraically – the following transformations:

1. Translate up by 2, then reflect about the x-axis, then translate left by 3.
2. Pull away from the y-axis by a factor of 3, then pull toward the x-axis by a factor of 2.

Exercise 3.8.14

What sequences of basic transformations discussed above produce these results?



Exercise 3.8.15

Describe – both geometrically and algebraically – a transformation that makes a 1×1 square into a 2×1 rectangle.

Exercise 3.8.16

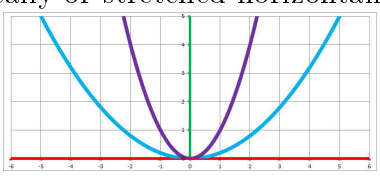
What transformations increase/decrease slopes of lines?

Exercise 3.8.17

Point out the inverses of each of the six transformations of the plane on the list.

Exercise 3.8.18

Has this parabola been shrunk vertically or stretched horizontally?



Example 3.8.19: more transformations as symmetries

Recall some of the transformations of the plane we saw in Chapter 2 when we discussed symmetry.

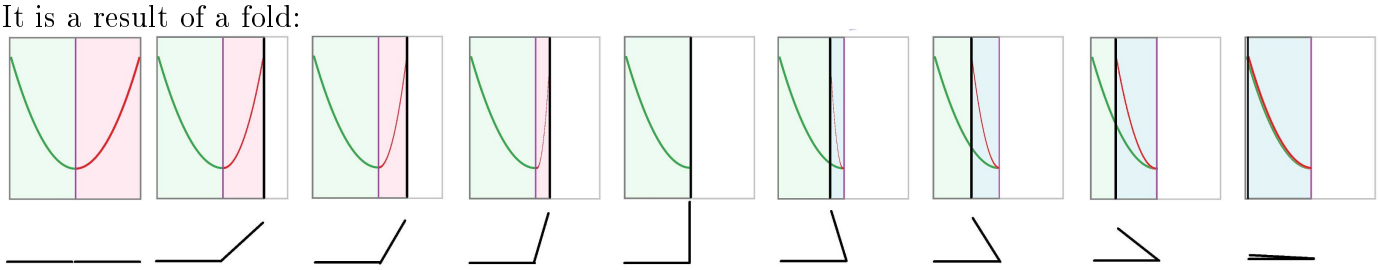
For example, the fact that the parabola's left branch is a mirror image of its right branch is revealed via a horizontal flip:

$$\boxed{(x, y) \mapsto (-x, y)}$$

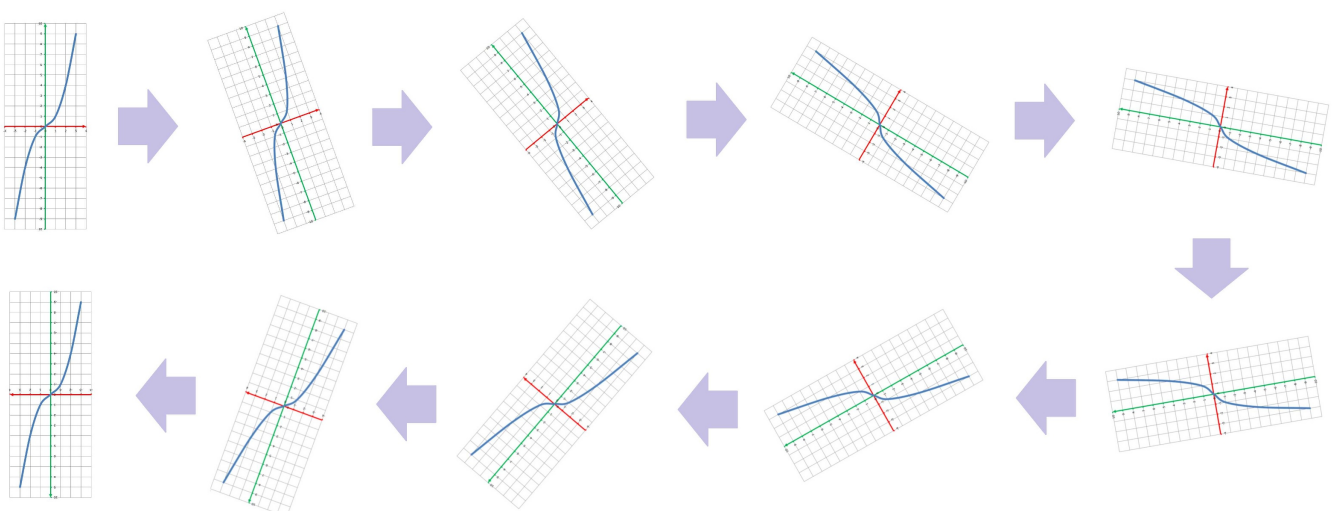
Generally, a subset A of the plane is *mirror symmetric* about the y-axis when we have:

$$(x, y) \text{ belongs to } A \implies (-x, y) \text{ belongs to } A\,.$$

It is a result of a fold:



Let's find a formula for a rotation 180 degrees around the origin, also called the *central symmetry*:



We achieve the same effect if we instead flip the plane about the y-axis and then about the x-axis (or

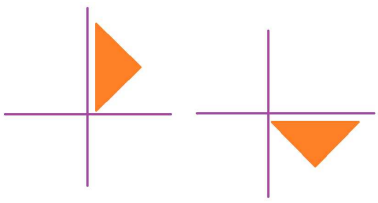
vice versa):

$$\boxed{(x,y) \mapsto (-x,y) \mapsto (-x,-y)}$$

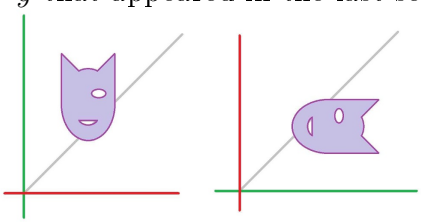
Then a subset A of the plane is *centrally symmetric* when we have:

$$(x,y) \text{ belongs to } A \implies (-x,-y) \text{ belongs to } A\,.$$

However, some transformations *cannot* be decomposed into a composition of those six basic transformations! Here is, for example, a 90-degree rotation:



Another is a flip about the line $x = y$ that appeared in the last section:



It is given by

$$(x,y) \mapsto (y,x)\,.$$

(Transformations of the plane are discussed in Chapter 4HD-3 and Chapter 5DE-2.)

Exercise 3.8.20

Find a formula for a 90-degree clockwise rotation.

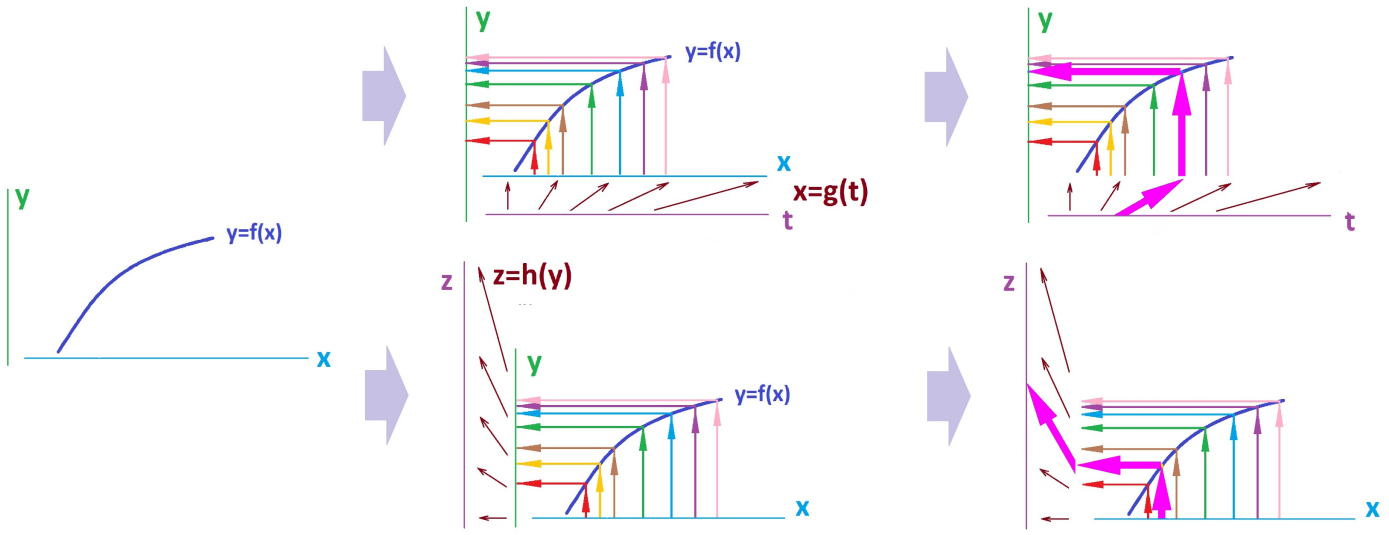
Exercise 3.8.21

What are the inverses of the transformations presented in this section?

3.9. Changing a variable transforms the graph of a function

Consider these three facts. First, a function is represented by its graph, which is a certain subset of the xy-plane. Second, a function is seen as a transformation of the real number line. Third, there are two prominent number lines on the xy-plane: the x-axis and the y-axis! So, the graph of a function "drawn" on the xy-plane is being transformed by functions of x or y.

Suppose we have the graph of a function $y = f(x)$. Suppose we form the compositions of f with other functions such as in the case when we switch from x, time in minutes, and y, location in inches, to t, time in seconds, or z, location in feet:



We know from the last section that we face transformations of the two axes. Even though we can follow the arrows and find the values of this composition, what does its *graph* look like?

We ask the following two questions:

1. What will happen to the graph of a function $y = f(x)$ if the x-axis is transformed by *another* function?
2. What will happen to the graph of a function $y = f(x)$ if the y-axis is transformed by *another* function?

A short answer to both is: The graph will transform into that of the *composition* of f with this function. The *order*, however, is different. It's "before f" vs. "after f":

1. $t \to \boxed{g} \to x \to \boxed{f} \to y$
2. $x \to \boxed{f} \to y \to \boxed{h} \to z$

This is why the answers to the two questions will be different.

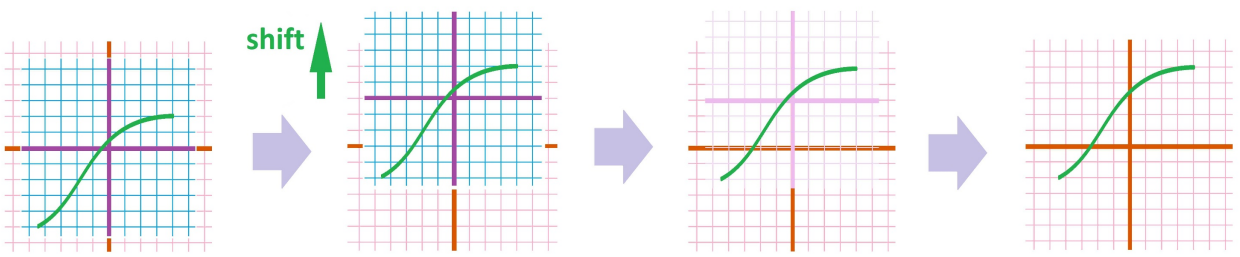
Warning!

Graphs aren't functions and functions aren't graphs; this is all about visualization.

If the problem is truly about units conversion, we'd better have the variables and the axes named differently. However, in this section, we pursue the idea of producing a new function on the *same* xy-plane.

For the six basic transformations discussed previously, there will be six rules governing the algebra of the functions affected by them.

We start with the *vertical shift*. Since the graph is drawn on the same piece of paper, it is shifted exactly the same way:



We see the original function $y = f(x)$ on the far left and the new function $y = F(x)$ on the far right. What is the formula of the new graph? If we take a point (x, y) on the graph of $y = x^2$ and shift up by 5, we have new point $(x, y + 5)$. It lies on the graph of $y = x^2 + 5$.

Theorem 3.9.1: Vertical Shift

If the graph $y = F(x)$ is the graph of $y = f(x)$ shifted s units up, then

$$F(x) = f(x) + s$$

and vice versa.

Proof.

In order to find the formula for the new function, we compare the following two facts about the two graphs:

- A point (x, y) lies on the graph of f, hence $f(x) = y$.
- The shifted point $(x, y + s)$ lies on the graph of the new function F, hence $F(x) = y + s$.

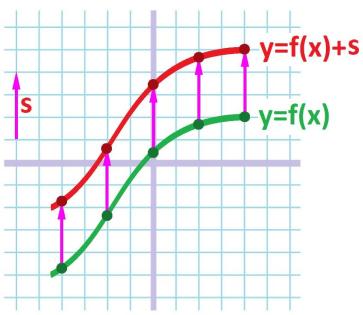
The latter was discovered earlier in this chapter. We substitute $y = f(x)$ and conclude:

$$F(x) = f(x) + s\,.$$

By "vice versa", we mean, of course, the *converse*:

▶ If a function g satisfies $F(x) = f(x) + s$, then its graph is the graph of f shifted s units up.

We can see how the green points are transferred to the red ones one by one:



The change is the same for all points. The most important conclusion is that the new function has the exact same *shape* of the graph as the old one.

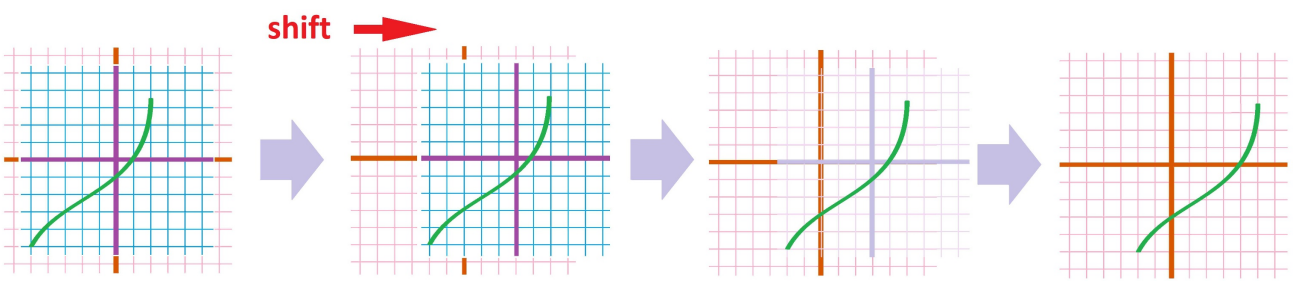
Example 3.9.2: formula for vertical shift

This is how easy it is to write the formula for a new function:

$$\begin{aligned} &\text{original:} && f(x) = && \frac{x^2 + \sqrt{x}}{x - 7} \\ &\text{shifted up by 3:} && F(x) = && \left(\frac{x^2 + \sqrt{x}}{x - 7}\right) \quad +3 \end{aligned}$$

We just put the original inside large parentheses and then add "+3" after.

Next is the *horizontal shift*:

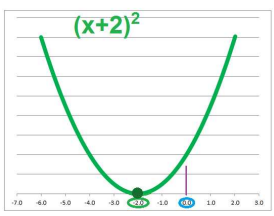


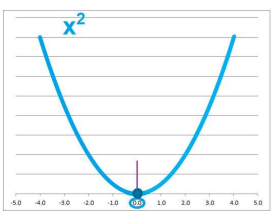
Example 3.9.3: parabola shifted

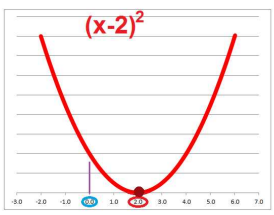
To guess the formula, let's take $f(x) = x^2$ and shift 2 units to the left and to the right:

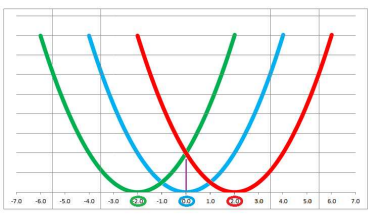
$$f(x+2) = (x+2)^2 \quad \text{and} \quad f(x-2) = (x-2)^2 .$$

These are the graphs of the three functions:









To confirm the match, what are the x-intercepts of these two new functions? Set either equal to 0 and solve:

$(x+2)^2 = 0$	$(x-2)^2 = 0$
$x+2 = 0$	$x-2 = 0$
$x = -2$	$x = 2$
left shift	right shift

The shift is in the direction *opposite* to the sign of what is added!

Theorem 3.9.4: Horizontal Shift

If the graph of $y = F(x)$ is the graph of $y = f(x)$ shifted s units to the right, then

$$F(x) = f(x-s)$$

and vice versa.

Proof.

In order to find the *formula for the new function*, we compare the following two facts about the two graphs:

- A point (x, y) lies on the graph of f, then $f(x) = y$.
- The shifted point $(x+s, y)$ lies on the graph of the new function F, hence $F(x+s) = y$.

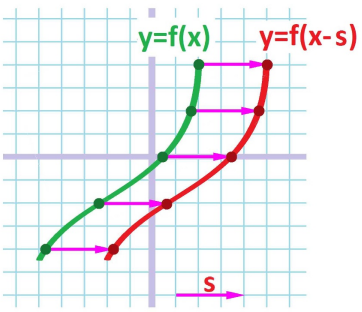
Substitute $y = f(x)$ and conclude:

$$F(x+s) = f(x) .$$

Therefore, the new function is the following:

$$F(x) = f(x-s) .$$

We can see how the green points are transferred to the red ones one by one:



The change is the same for all points. Once again, the new function has the exact same *shape* of the graph as the old one.

Example 3.9.5: formula for horizontal shift

This is how easy it is to write the formula for a new function:

$$\text{original:} \qquad f(x) = \frac{x^2 + \sqrt{x}}{x - 7}$$

$$\text{shifted right by 3:} \qquad F(x) = \frac{(x-3)^2 + \sqrt{(x-3)}}{(x-3) - 7}$$

We just replace each x in the original with "$(x-3)$".

Example 3.9.6: two shifts compared

Recall the algorithmic interpretation of these two types of shifts:

vertical shift, up 3 : $x \to \boxed{f} \to y \to \boxed{\text{add 3}} \to y$

horizontal shift, right 3 : $x \to \boxed{\text{subtract 3}} \to x \to \boxed{f} \to y$

Example 3.9.7: two shifts combined

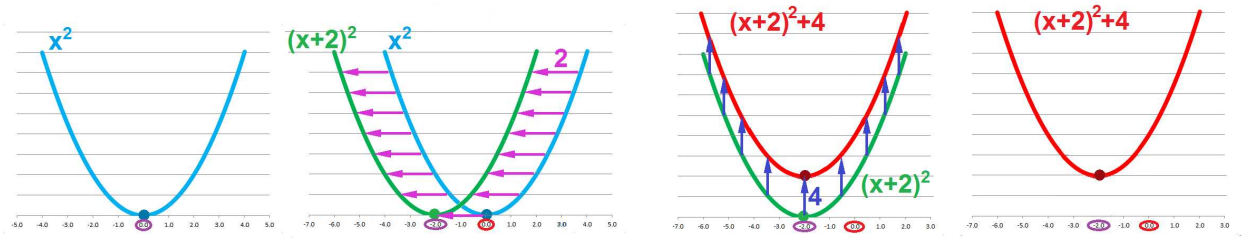
What about "combination" of these two types of transformations? Let

$$Q(x) = (x+2)^2 + 4\,.$$

What is the graph? We can plot it point by point, but as a preview of things to come, let's decompose the function:

$$x \to \boxed{x+2} \to u \to \boxed{u^2} \to z \to \boxed{z+4} \to y$$

This is a 2-unit leftward shift followed by a 4-unit upward shift:



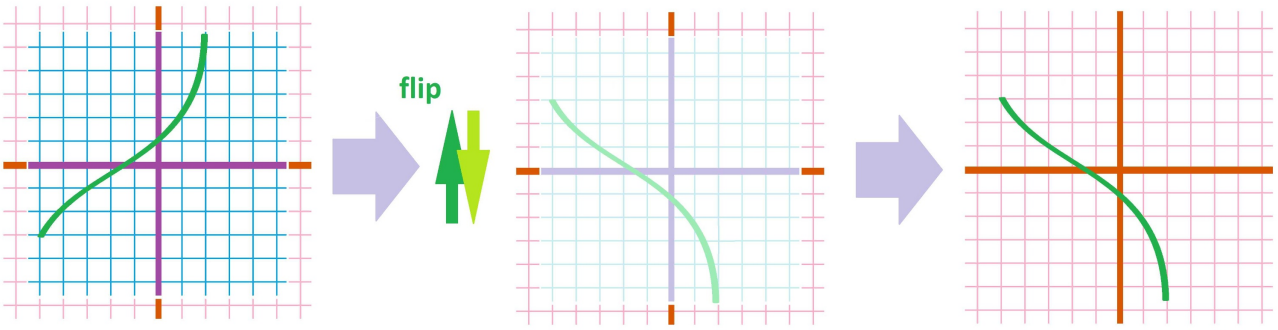
Exercise 3.9.8

What happens to the x- and y-intercepts of a function under vertical and horizontal shifts?

Exercise 3.9.9

What happens to the domain and the range of a function under vertical and horizontal shifts?

Next is the *vertical flip*:



If we take a point (x, y) on the graph of $y = x^2$ and flip about the x-axis, we have new point $(x, -y)$. It lies on the graph of $y = -x^2$.

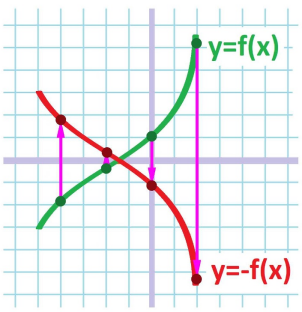
Theorem 3.9.10: Vertical Flip

If the graph of $y = F(x)$ is the graph of $y = f(x)$ flipped vertically, then

$$F(x) = -f(x)$$

and vice versa.

We can see how the green points are transferred to the red ones one by one:



The distance from every new point to the x-axis is that same as that of the original point but on the other side. Again, the new function has the exact same shape of the graph as the old one.

Exercise 3.9.11

Prove the theorem.

Example 3.9.12: formula for vertical flip

This is how easy it is to write the formula for a new function:

$$\begin{aligned} &\text{original:} && f(x) = && \frac{x^2 + \sqrt{x}}{x - 7} \\ &\text{flipped vertically:} && F(x) = && -\left(\frac{x^2 + \sqrt{x}}{x - 7}\right) \end{aligned}$$

We just put the original inside large parentheses and then add "$-$" in front.

Example 3.9.13: order might matter

Let

$$Q(x) = -f(x) + 3\,.$$

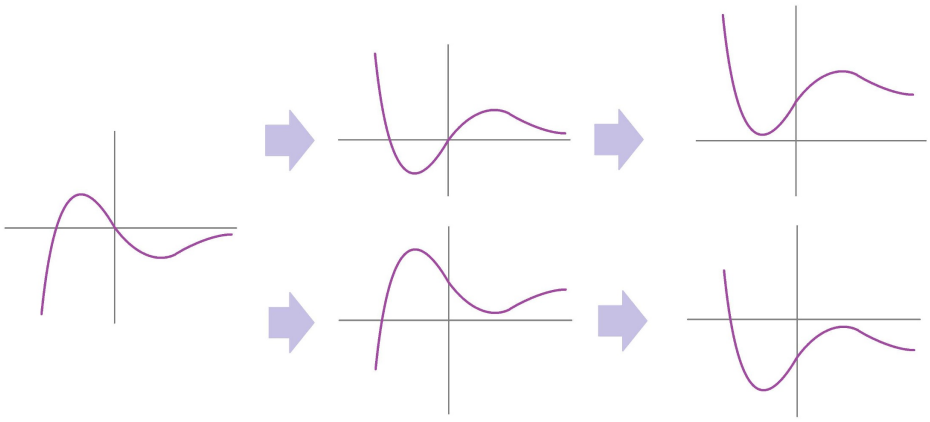
How do we get the graph of h? Let's decompose and provide matching transformations:

$$x \longmapsto f(x) \underbrace{\longmapsto}_{\text{vertical flip}} -f(x) \underbrace{\longmapsto}_{\text{vertical shift}} -f(x) + 3$$

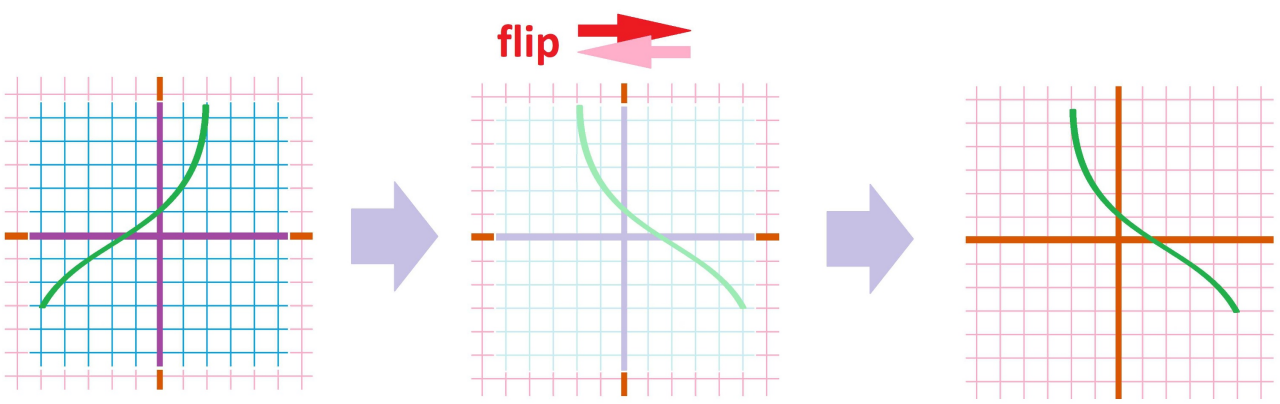
The order matters! Let's interchange these two:

$$x \longmapsto f(x) \underbrace{\longmapsto}_{\text{vertical shift}} f(x) + 3 \underbrace{\longmapsto}_{\text{vertical flip}} -(f(x) + 3)$$

The graphs are also different:



Next is the *horizontal flip*:



If we take a point (x, y) on the graph of $y = x^2$ and flip about the y-axis, we have new point $(-x, y)$. It lies on the graph of $y = (-x)^2$. It's the same graph because of the mirror symmetry.

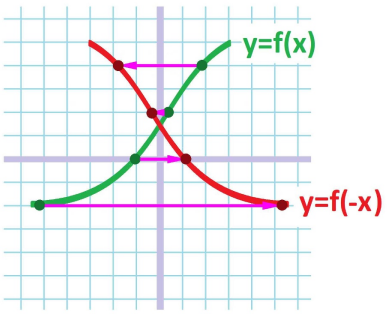
Theorem 3.9.14: Horizontal Flip

If the graph of $y = F(x)$ is the graph of $y = f(x)$ flipped horizontally, then

$$F(x) = f(-x)$$

and vice versa.

We can see how the green points are transferred to the red ones one by one:



The distance from every new point to the y-axis is that same as that of the original point but on the other side. Once again, the new function has the exact same shape of the graph as the old one. In fact, all four transformations are *rigid motions*.

Exercise 3.9.15

Prove the theorem.

Example 3.9.16: formula for horizontal flip

This is how easy it is to write the formula for a new function:

$$\text{original:} \quad f(x) = \frac{x^2 + \sqrt{x}}{x - 7}$$

$$\text{flipped horizontally:} \quad F(x) = \frac{(-x)^2 + \sqrt{(-x)}}{(-x) - 7}$$

We just replace each x in the original with "$(-x)$".

Exercise 3.9.17

Does the order matter when the horizontal flip is combined with a horizontal shift? What about the horizontal flip with a *vertical* shift?

Example 3.9.18: two flips compared

These are the diagrams:

vertical flip: $x \to \boxed{f} \to y \to \boxed{\text{multiply by } -1} \to y$

horizontal flip: $x \to \boxed{\text{multiply by } -1} \to x \to \boxed{f} \to y$

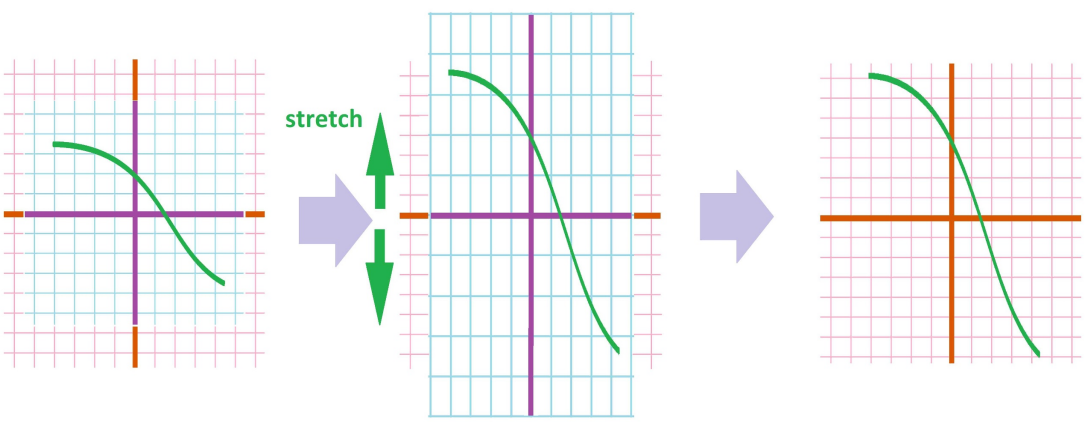
Exercise 3.9.19

What happens to the x- and y-intercepts of a function under vertical and horizontal flips?

Exercise 3.9.20

What happens to the domain and the range of a function under vertical and horizontal flips?

Next is the *vertical stretch*:



If we take a point (x, y) on the graph of $y = x^2$ and stretch vertically by 3, we have new point $(x, 3y)$. It lies on the graph of $y = 3x^2$.

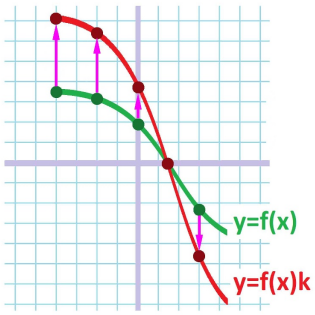
Theorem 3.9.21: Vertical Stretch

If the graph of $y = F(x)$ is the graph of $y = f(x)$ stretched vertically by a factor of $k > 0$, then

$$F(x) = kf(x)$$

and vice versa.

We can see how the green points are transferred to the red ones one by one:



In case of $k = 2$, the distance from every new point to the original point is the same as from the original to the x-axis. It is clear that the new function has a somewhat different shape of the graph than the old one.

Exercise 3.9.22

Prove the theorem.

Example 3.9.23: formula for vertical stretch

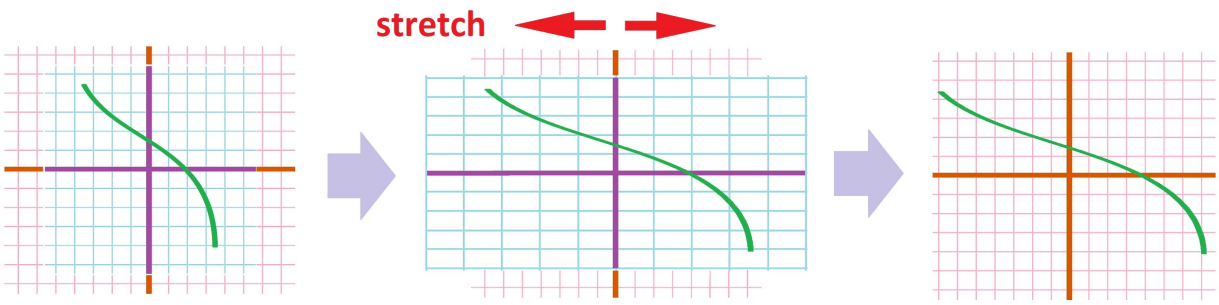
This is how easy it is to write the formula for a new function:

$$\text{original:} \qquad f(x) = \frac{x^2 + \sqrt{x}}{x - 7}$$

$$\text{stretched vertically by 5:} \qquad F(x) = 5 \cdot \left(\frac{x^2 + \sqrt{x}}{x - 7} \right)$$

We just put the original inside large parentheses and then add "5" in front.

Next is the *horizontal stretch*:



We can guess that $y = x^2$ stretched horizontally by 3 becomes $y = (x/3)^2$.

Theorem 3.9.24: Horizontal Stretch

If graph $y = F(x)$ is the graph of $y = f(x)$ stretched by the factor $k > 0$ horizontally, then

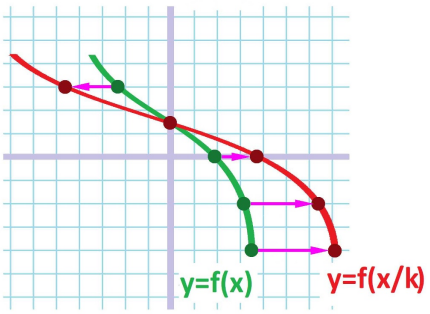
$$F(x) = f(x/k)$$

and vice versa.

Exercise 3.9.25

Prove the theorem.

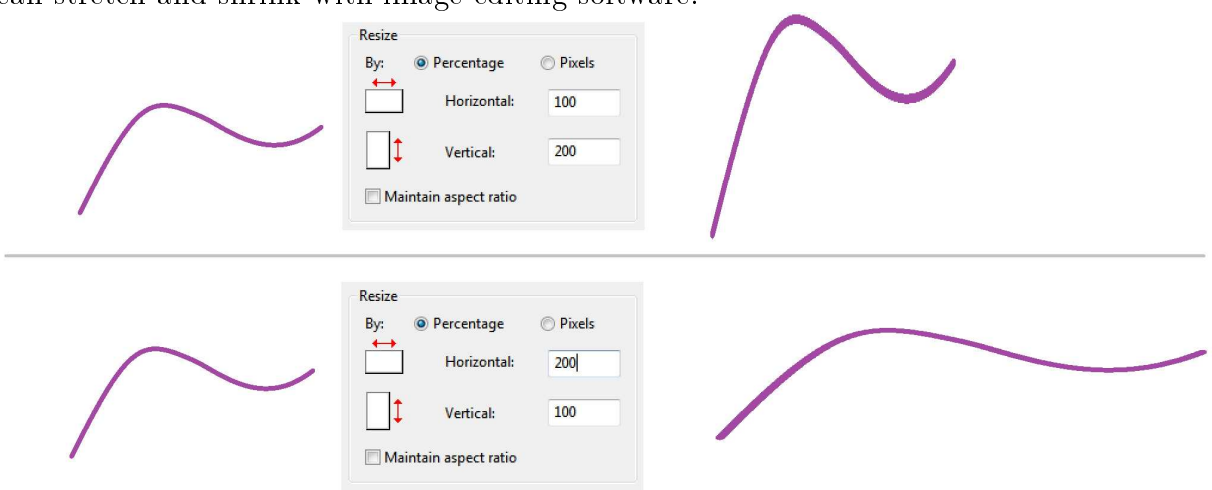
We can see how the green points are transferred to the red ones one by one:



In case of $k = 2$, the distance from every new point to the original point is the same as from the original to the y-axis. It is clear that the new function has a somewhat different shape of the graph than the old one. The last two transformations aren't rigid motions.

Example 3.9.26: transformations with computer graphics

One can stretch and shrink with image editing software:



Example 3.9.27: formula for horizontal stretch

This is how easy it is to write the formula for a new function:

$$\text{original:} \qquad f(x) = \frac{x^2 + \sqrt{x}}{x - 7}$$

$$\text{stretched horizontally by 5 :} \qquad F(x) = \frac{(x/5)^2 + \sqrt{(x/5)}}{(x/5) - 7}$$

We just replace each x in the original with "$(x/5)$".

Example 3.9.28: two stretches compared

Now the algorithmic interpretation of these operations:

$$\text{vertical stretch by 3 :} \qquad x \to \boxed{f} \to y \to \boxed{\text{multiply by 3}} \to z$$

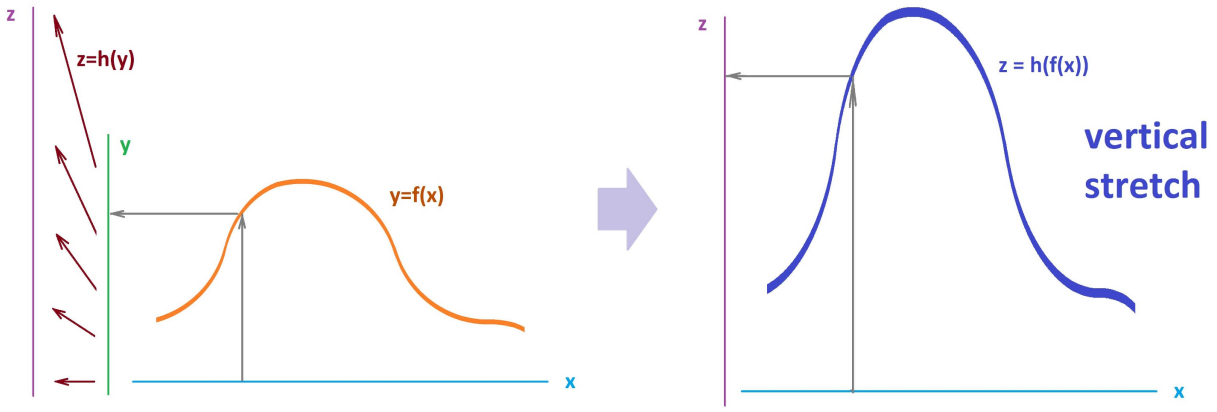
$$\text{horizontal stretch by 3 :} \qquad t \to \boxed{\text{divide by 3}} \to x \to \boxed{f} \to y$$

The theorem below summarizes our analysis.

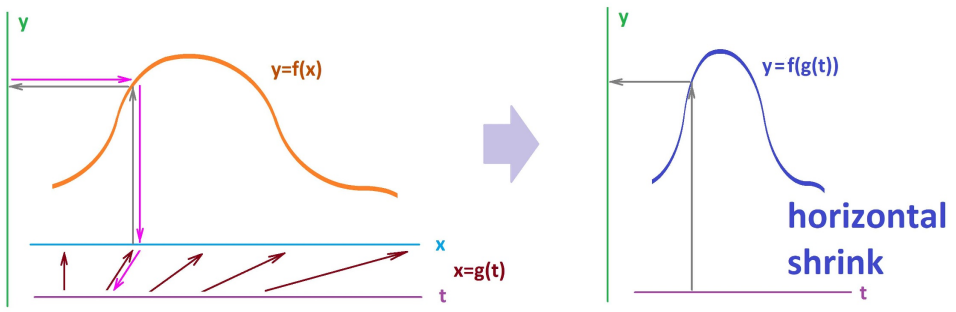
Theorem 3.9.29: Transformations as Compositions

1. *A vertical transformation of the graph of a function $y = f(x)$ results from its composition with a function that follows f, i.e., $h \circ f$.*
2. *A horizontal transformation of the graph of a function f results from its composition with a function that precedes f, i.e., $f \circ g$.*

For part 1, this is what $z = 2y$ does to the graph of $y = f(x)$:



For part 2, this is what $x = 2t$ does to the graph of $y = f(x)$:



As we can see, such a transformation of an axis is a *change of variables* (or units). Every function f has *two*: input and output. Then:

- In the former case, we change the output variable: from y to $z = h(y)$.
- In the latter case, we change the input variable: from x to $t = g^{-1}(x)$.

This difference is the reason why the effect on the graph of f is so different.

Exercise 3.9.30

What happens to the x- and y-intercepts of a function under vertical and horizontal stretches?

Exercise 3.9.31

What happens to the domain and the range of a function under vertical and horizontal stretches?

Exercise 3.9.32

(a) How do these transformations affect the monotonicity of a function? (b) Investigate the monotonicity of quadratic functions.

Exercise 3.9.33

By transforming the graph of $y = x^2$, plot the graphs of the functions: (a) $y = \sqrt{x}$ and (b) $y = \sqrt{x+3}$.

In conclusion, composing a function with a *linear function* – before or after – will transform its graph in these six ways: shift, flip, and stretch, vertical or horizontal.

Summary of the rules:

	vertical, y		horizontal, x	
shift by s :	$y = f(x)$	$+s$	$y = f(x$	$-s)$
flip :	$y = f(x)$	$\cdot(-1)$	$y = f(x$	$\cdot(-1))$
stretch by k :	$y = f(x)$	$\cdot k$	$y = f(x$	$/k)$

In the "horizontal" column, we have the formulas of the *inverses*:

- Add s vs. subtract s.
- Multiply by (-1) vs. multiply by (-1).
- Multiply by $k > 0$ vs. divide by k.

Inverses appear every time we solve an equation to find the formula for the new function, F, after a change of the input variable (i.e., a horizontal transformation) of a function f. For example, if we have $F(x+s) = f(x)$ after a horizontal shift, we substitute $u = x + s$ and find x in terms of u, i.e., $x = u - s$, resulting in $F(u) = f(u - s)$. After an optional renaming, u for x, the formula takes its final form, $F(x) = f(x - s)$.

Example 3.9.34: order of operations vs. order of transformations

There is a bit more complexity here: The order in which the transformations are carried out follows the direction *away* from f. Suppose we have this function:

$$\ldots \to \boxed{\text{add } 5} \to \boxed{\text{multiply by } -1} \to \underset{\text{horizontal}}{x} \to \boxed{f} \to \underset{\text{vertical}}{y} \to \boxed{\text{subtract } 3} \to \boxed{\text{multiply } 5} \to \ldots$$

Then, this is what these functions do to the graph of f:

- shift down by 3 and then stretch by 5 (left to right);
- flip horizontally and then shift left by 5 (right to left).

3.10. The graph of a quadratic polynomial is a parabola

In this section, we will concentrate on a *specific* class of functions.

We have called the graph of $y = x^2$ a "parabola". Are there others?

The graph of every quadratic polynomial is a parabola.

This is what it means:

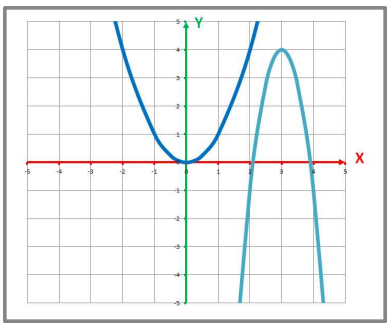
▶ The graph of any quadratic polynomial can be acquired from *the* parabola of $f(x) = x^2$ via our six basic transformations (or even fewer).

We are about to carry out this plan:

$$y = x^2 \quad \xrightarrow{\text{transformations}} \quad y = ax^2 + bx + c, \ a \neq 0$$

Example 3.10.1: transforming parabola

Suppose the graph of g is given on the right. We need to transform the graph of $y = x^2$ into this parabola:



We start with a comparison.

- The most obvious feature is that the right one *opens down*. We will, therefore, have to do a vertical flip.
- The second most prominent feature is that the right one is *slimmer*. We will have to do a vertical stretch or a horizontal shrink.
- The last one is the *location*: The vertex of the parabola is away from the origin. We will have to do both vertical and horizontal shifts.

Note that a horizontal flip is useless here because the parabola has a vertical mirror symmetry. Also note that we pick a vertical stretch over a horizontal shrink as the simpler one.

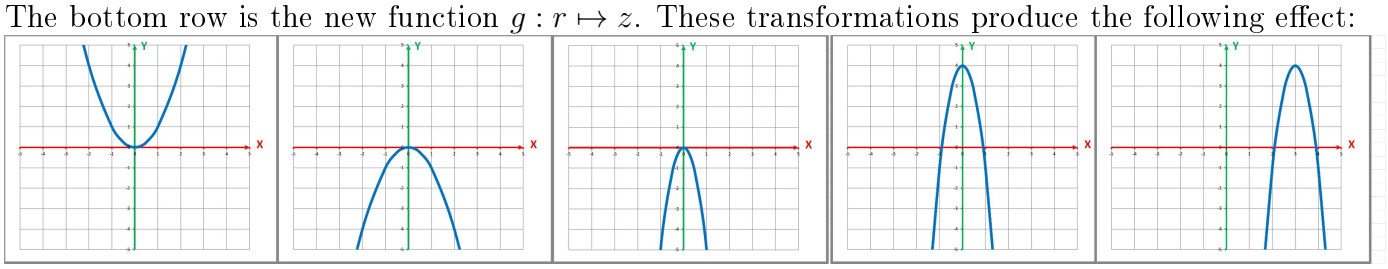
We now turn to the actual algebra. What is the order of the operations that we have outlined? We do all of the vertical first, then horizontal:

$$\begin{array}{rll}
\text{original:} & x^2 & \\
\text{vertical flip:} & x^2 \cdot (-1) & = -x^2 \\
\text{vertical stretch:} & x^2 \cdot (-1) \cdot 5 & = -5x^2 \\
\text{vertical shift:} & x^2 \cdot (-1) \cdot 5 + 4 & = -5x^2 + 4 \\
\text{horizontal shift:} & (x-3)^2 \cdot (-1) \cdot 5 + 4 & = -5(x-3)^2 + 4
\end{array}$$

This algebra comes from the following sequence of compositions:

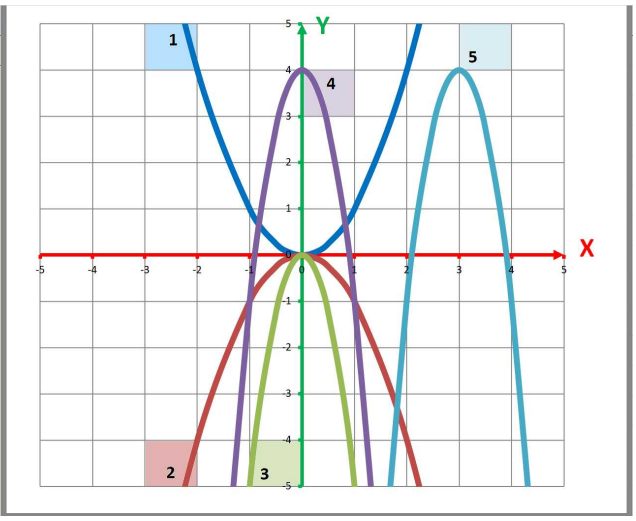
$$\begin{array}{ccccccccc}
x & \to & x^2 = y & \to & -y = u & \to & 5u = v & \to & v + 4 = w \\
\downarrow & & & & & & & & \| \\
x + 3 = r & \longrightarrow & \longrightarrow & \longrightarrow & \longrightarrow & \longrightarrow & \longrightarrow & \longrightarrow & w = z
\end{array}$$

The bottom row is the new function $g : r \mapsto z$. These transformations produce the following effect:



One can also see how the data is changing as we progress through the sequence:

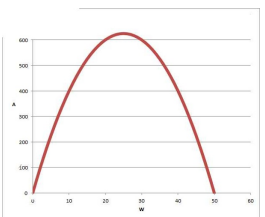
	1	2	3	4		5
x	y	u=-y	v=5u	w=v+4	r=x+3	z=w
-5.0	25.0	-25.0	-125.0	-121.0	-2.0	-121.0
-4.0	16.0	-16.0	-80.0	-76.0	-1.0	-76.0
-3.0	9.0	-9.0	-45.0	-41.0	0.0	-41.0
-2.0	4.0	-4.0	-20.0	-16.0	1.0	-16.0
-1.0	1.0	-1.0	-5.0	-1.0	2.0	-1.0
-0.5	0.3	-0.3	-1.3	2.8	2.5	2.8
-0.3	0.1	-0.1	-0.3	3.7	2.8	3.7
0.0	0.0	0.0	0.0	4.0	3.0	4.0
0.3	0.1	-0.1	-0.3	3.7	3.3	3.7
0.5	0.3	-0.3	-1.3	2.8	3.5	2.8
1.0	1.0	-1.0	-5.0	-1.0	4.0	-1.0
2.0	4.0	-4.0	-20.0	-16.0	5.0	-16.0
3.0	9.0	-9.0	-45.0	-41.0	6.0	-41.0
4.0	16.0	-16.0	-80.0	-76.0	7.0	-76.0
5.0	25.0	-25.0	-125.0	-121.0	8.0	-121.0



The first two columns form the original function and the final function is in the last two columns.

But can we go *backwards* and find how to represent a quadratic function given to us, as a formula, as a result of such a sequence of transformations?

To begin with, let's learn how to find the vertex of a parabola:



Suppose the polynomial is given by a formula, its *standard representation*:

$$f(x) = ax^2 + bx + c,\ a \neq 0\,.$$

Judging by the example, we need to morph it into the following form:

$$f(x) = a(x-h)^2 + k$$

While a contains information about the stretch/shrink and the flip of the graph, h and k are the shifts. In fact, the point (h, k) is the *vertex* of the parabola.

We will use the *complete square formula*:

$$(u+v)^2 = u^2 + 2uv + v^2$$

Example 3.10.2: completing a square

Let's show how this is done algebraically. Suppose

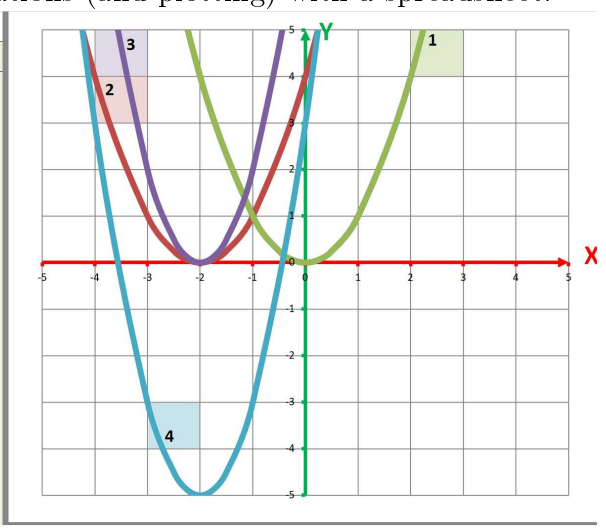
$$f(x) = 2x^2 + 8x + 3\,.$$

We manipulate the formula towards our goal:

$$\begin{aligned}
f(x) &= 2x^2+8x+3 && \text{Start with the original.}\\
&= (2x^2+8x)+3 && \text{Bring together the two terms with } x.\\
&= 2(x^2+4x)+3 && \text{Factor, so that you have } x^2.\\
& && \text{The other term is half the coefficient of } x,\ 4/2=2.\\
&= 2(x^2+4x+2^2-2^2)+3 && \text{Add the missing term, } 2^2, \text{ of the complete square.}\\
&= 2(x^2+4x+2^2)-2\cdot 2^2+3 && \text{Pull out the extra term.}\\
&= 2(x+2)^2-8+3 && \text{Complete the square.}\\
&= 2(x+2)^2-5 && \text{Acquire the final form.}
\end{aligned}$$

Reading from the inside out: Shift left by 2, stretch vertically by 2, shift down by 5. The vertex is at $(-2,-5)$. To confirm, we carry out these computations (and plotting) with a spreadsheet:

	1		2	3	4
x	y	t=x-2	y	u=2y	v=u-5
-5.0	25.0	-7.0	25.0	50.0	45.0
-4.0	16.0	-6.0	16.0	32.0	27.0
-3.0	9.0	-5.0	9.0	18.0	13.0
-2.0	4.0	-4.0	4.0	8.0	3.0
-1.0	1.0	-3.0	1.0	2.0	-3.0
-0.5	0.3	-2.5	0.3	0.5	-4.5
-0.3	0.1	-2.3	0.1	0.1	-4.9
0.0	0.0	-2.0	0.0	0.0	-5.0
0.3	0.1	-1.8	0.1	0.1	-4.9
0.5	0.3	-1.5	0.3	0.5	-4.5
1.0	1.0	-1.0	1.0	2.0	-3.0
2.0	4.0	0.0	4.0	8.0	3.0
3.0	9.0	1.0	9.0	18.0	13.0
4.0	16.0	2.0	16.0	32.0	27.0
5.0	25.0	3.0	25.0	50.0	45.0



Exercise 3.10.3

How can you transform *any* parabola into *any* other parabola?

A quadratic polynomial may be called a "complete square" if it can be put in this form:

$$f(x) = (x-h)^2 = x^2+2xh+h^2\,,$$

for some number h. It is easy to plot; just shift the original h units right. This idea has a broader applicability, but the challenge is to *recognize* complete squares in (or extract from) quadratic polynomials.

Theorem 3.10.4: Vertex Form of Quadratic Polynomial

Any quadratic polynomial,

$$f(x) = ax^2+bx+c,\ a\neq 0\,,$$

can be represented as an "incomplete square" or a "vertex form":

$$f(x) = a(x-h)^2+k$$

where h, k are the following numbers:

$$h = -\frac{b}{2a},\quad k = c-ah^2$$

Proof.

All we need is to find these parameters: h and k. First, we set the two representations of the same function equal to each other:

$$f(x) \quad = ax^2 + bx + c \quad = a(x-h)^2 + k\,.$$

We then expand the latter and align its terms with the former:

$$\begin{array}{rccccccc} & a & \cdot x^2 & + & b & \cdot x & +c \\ = & a & \cdot x^2 & + & 2ah & \cdot x & +(ah^2+k)\,. \end{array}$$

Then we match up the coefficients of the terms:

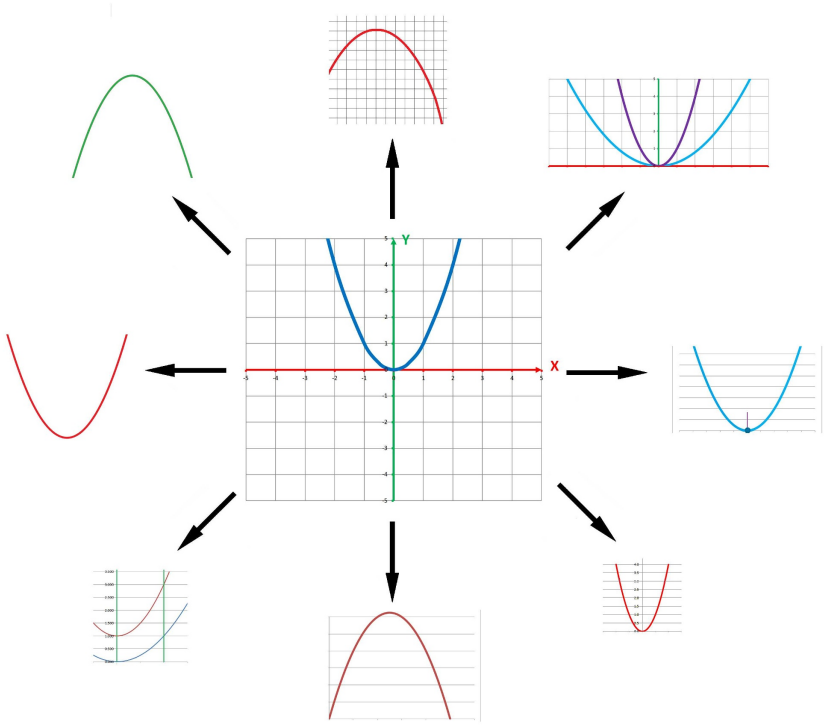
$$b = 2ah,\ \ c = ah^2 + k\,.$$

And finally we solve for h and k.

Exercise 3.10.5

Justify matching the coefficients in the proof.

Consequently, the graph of f can be acquired from the graph of $y = x^2$ via the six transformations, or just four if we exclude the vertical flip and the horizontal shrink. We can say that *there is only one parabola*:



Exercise 3.10.6

How can the formula for h above be extracted from the Quadratic Formula?

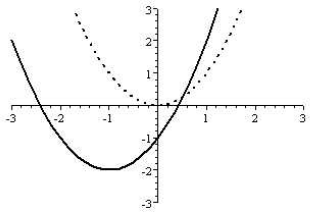
In summary, we have the following:

Definition 3.10.7: parabola

A *parabola* is the graph of a quadratic polynomial.

Exercise 3.10.8

The graphs below are parabolas. One is $y = x^2$. What is the other?



In conclusion, composing a function f with another function,

$$h \circ f \quad \text{or} \quad f \circ g\,,$$

transforms its graph vertically or horizontally, respectively.

Theorem 3.10.9: Transformations of Graphs

The graph of $y = F(x)$ is the graph of $y = f(x)$ transformed as indicated if and only if its formula is provided in the table below:

	vertically		*horizontally*	
shifted by s :	$y = f(x)$	$+s$	$y = f(x$	$-s)$
flipped :	$y = f(x)$	$\cdot(-1)$	$y = f(x$	$\cdot(-1))$
stretched by k :	$y = f(x)$	$\cdot k$	$y = f(x$	$/k)$

Proof.

The transformations of the y-axis come from a function applied *after* f:

vertical:		$z = h(y)$,	substitute $y = f(x)$,	$z = h(f(x))$
y	$\xrightarrow{\text{shift by } k}$	$z = y + k$	$\Longrightarrow$	$z = f(x) + k$
y	$\xrightarrow{\text{flip}}$	$z = -y$	$\Longrightarrow$	$z = -f(x)$
y	$\xrightarrow{\text{stretch by } k}$	$z = y \cdot k$	$\Longrightarrow$	$z = f(x)k$

The transformations of the x-axis come from a function applied *before* f:

horizontal:		$x = g(t)$	substituted into $y = f(x)$,	$y = f(g(t))$
t	$\xrightarrow{\text{shift by } k}$	$x = t - k$	$\Longrightarrow$	$y = f(t - k)$
t	$\xrightarrow{\text{flip}}$	$x = -t$	$\Longrightarrow$	$y = f(-t)$
t	$\xrightarrow{\text{stretch by } k}$	$x = t/k$	$\Longrightarrow$	$y = f(t/k)$

Example 3.10.10: vertical stretching vs. horizontal shrinking

It might seem that vertical stretching is somehow equivalent to horizontal shrinking. We would never make that mistake when dealing with a real-life object, but we might be confused by an *infinite* shape, such as a line or a parabola... Consider

$$f(x) = x\,.$$

Stretched by a factor 2 vertically, it becomes:

$$F(x) = 2x\,.$$

Shrunk by a factor 2 horizontally, it becomes:

$$Q(x) = x \Big/ \frac{1}{2} = 2x\,.$$

This *is* the same function!

However, if we do the same to $f(x) = x^2$, we discover that

$$2x^2 \neq (2x)^2 .$$

This is *not* the same function!

Let's try the horizontal stretch by $\sqrt{2}$. Then

$$H(x) = (\sqrt{2}x)^2 .$$

Now there is a *match*:

$$\left(\sqrt{2}x\right)^2 = 2x^2 .$$

We don't expect such a match for most functions; just try $f(x) = 1$!

Chapter 4: The main classes of functions

Contents

4.1 The simplest functions 291
4.2 Monotonicity and the extreme values of functions 301
4.3 Functions with symmetries 310
4.4 Quadratic polynomials 321
4.5 The polynomial functions 328
4.6 The rational functions 341
4.7 The root functions 353
4.8 The exponential functions 360
4.9 The logarithmic functions 373
4.10 The trigonometric functions 378

4.1. The simplest functions

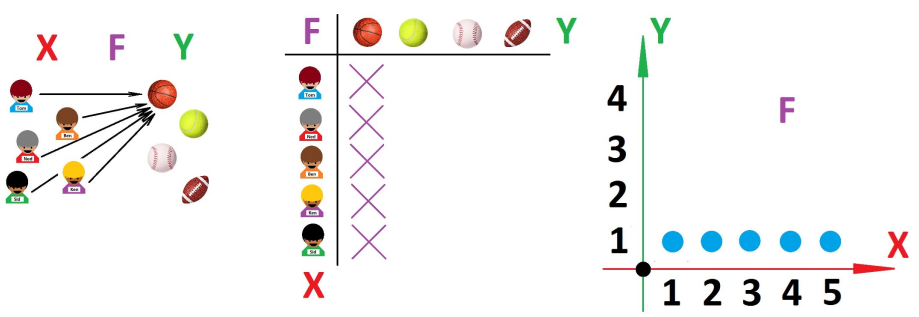
In this chapter, we will study many specific functions as well as some broad categories of functions. We start with the former.

Even in the most general situation – nothing but sets – there are always two functions that are very simple.

Let's turn to the example of the two sets we considered in Chapter 2:

- X is the five boys; and
- Y is the four balls.

Now, what if *all* boys prefer basketball? Then the "preference function", F, cannot be simpler: All of its values are equal and all the arrows point to the basketball:



The table of this function F is also very simple: All crosses are in the same column. The graph is just as simple: All dots are on the same horizontal line.

The value of $y = F(x)$ doesn't vary as x varies; it is *constant*. The following concept will be routinely used.

Definition 4.1.1: constant function

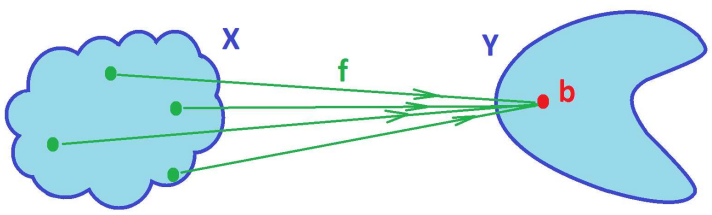
Suppose sets X and Y are given. A function $f : X \to Y$ is called a *constant function* if, for some specified element b of Y, we set:

$$f(x) = b \quad \texttt{FOR EACH } x$$

The process is identical for every input:

$$x \to \boxed{\text{choose } 3} \to y$$

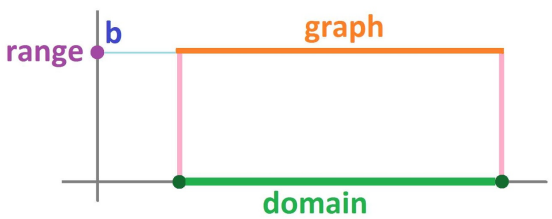
In the generic illustration below, all arrows converge on a single output:

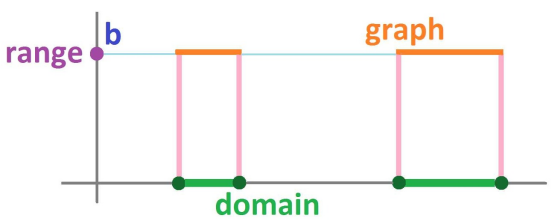


Exercise 4.1.2

What is the range of a constant function?

The graph of a constant *numerical* function is a horizontal line:





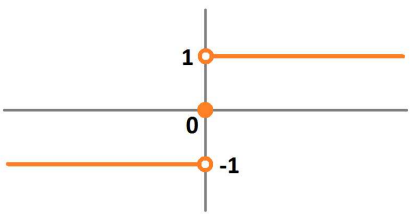
Of course, if the domain is disconnected, then so is the graph.

Example 4.1.3: step function

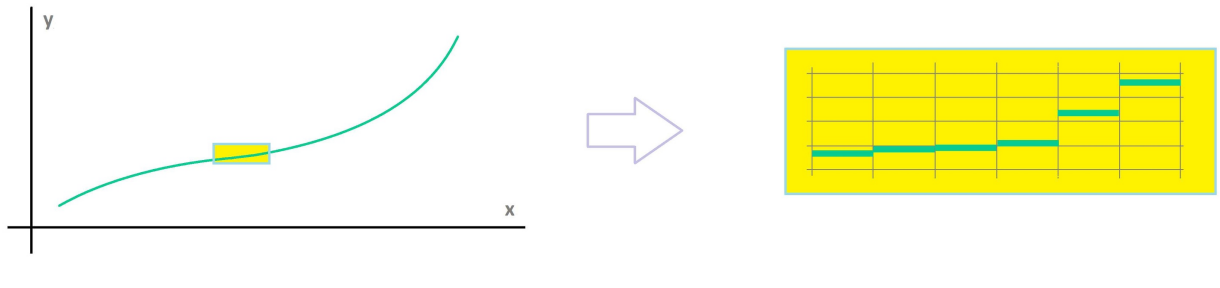
Constant functions are convenient building blocks for more complex functions. Here is a familiar example of how we build from three constant functions a single "piecewise constant" function:

$$y = \operatorname{sign}(x): \quad x \to \quad x \quad \begin{array}{lll} \nearrow \text{ if } x>0 & \boxed{\text{choose } 1} & \to y \searrow \\ \to \text{ if } x=0 & \boxed{\text{choose } 0} & \to y \to \quad y \to y \\ \searrow \text{ if } x<0 & \boxed{\text{choose } -1} & \to y \nearrow \end{array}$$

It is also often called a *step function*:



In fact, if we zoom in on a curve, we might discover that this is the graph of such a function:



Exercise 4.1.4

Build a few such functions from sequences.

What can we say about *compositions* of this special function, f, with another, g? First, consider this diagram:

$$\begin{array}{ccccc} X & \xrightarrow{\quad f \quad} & Y & \xrightarrow{\quad g \quad} & Z \\ \hline x_1 & & y_1 & \to & z_1 \\ & \searrow & & & \\ x_2 & \to & y_2 & \to & z_2 \\ & \nearrow & & & \\ x_3 & & y_3 & \to & z_3 \end{array}$$

What kind of function is $g \circ f$? And here is another one:

$$\begin{array}{ccccc} Z & \xrightarrow{\quad g \quad} & X & \xrightarrow{\quad f \quad} & Y \\ \hline z_1 & \to & x_1 & & y_1 \\ & & & \searrow & \\ z_2 & \to & x_2 & \to & y_2 \\ & & & \nearrow & \\ z_3 & \to & x_3 & & y_3 \end{array}$$

What kind of function is $f \circ g$? So, whether the other function comes after or before a constant function, the result is the same; this is our conclusion.

Theorem 4.1.5: Compositions with Constant Function

The composition of any function with a constant function is a constant function.

Proof.

This is the algebra for case 1:

$$x \mapsto f(x) = b \implies x \mapsto g(f(x)) = g(b)\,.$$

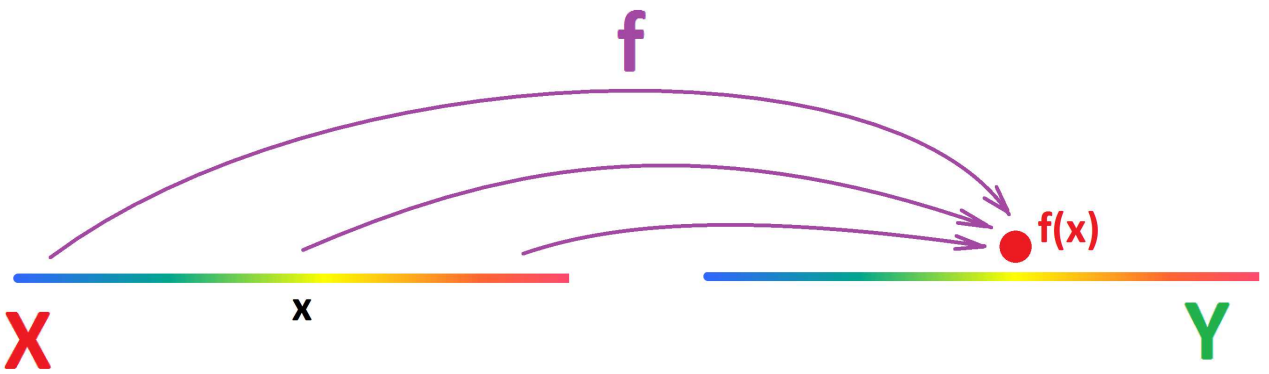
This is the algebra for case 2:

$$y \mapsto g(y) = x \implies y \mapsto f(g(y)) = b\,.$$

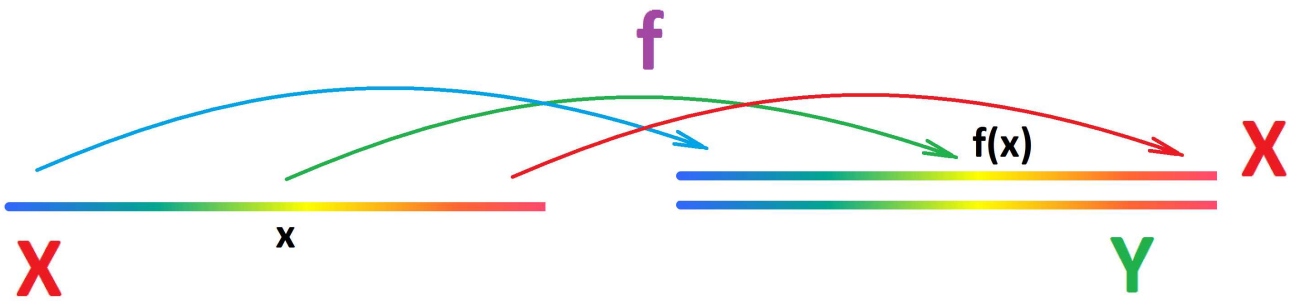
Exercise 4.1.6

Give an example of two non-constant functions the composition of which is constant.

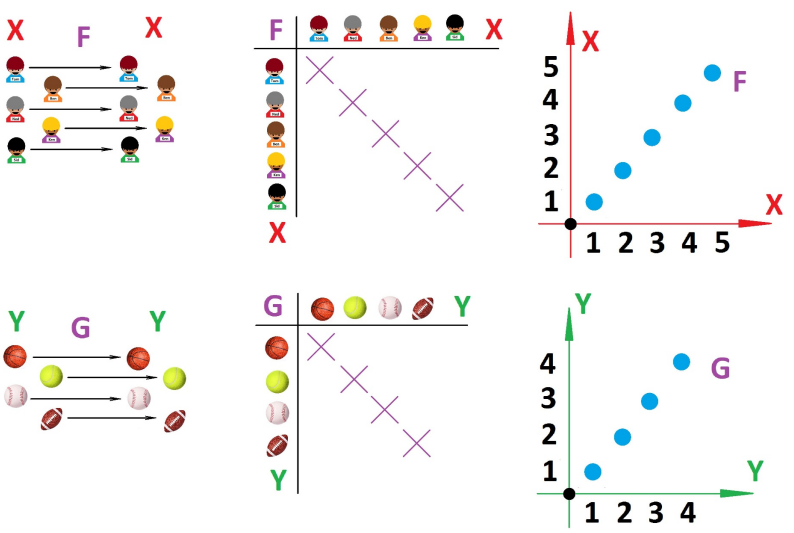
As a *transformation* of the line, the constant function is extreme; it "crushes" (or collapses) the whole line into a single point:



At the other end of the spectrum is another extreme transformation; it "does nothing" to the line:



So, for our set X of boys, we have a special function G from X to X (and another from Y to Y for the balls); each arrow comes back to the boy (or ball) it starts from:



The table of this function G is also very simple: All crosses are on the diagonal, and the graph has all dots on the diagonal. The output of G is *identical* to the input.

Definition 4.1.7: identity function

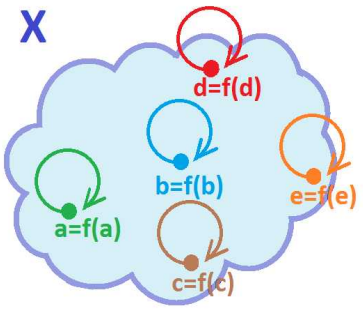
Suppose one set X is given. The *identity function* $I : X \to X$ is given by the following:

$$I(x) = x \quad \texttt{FOR EACH } x$$

The process is identical for every input:

$$x \to \boxed{\text{pass it}} \to y$$

In the generic illustration below, every arrow circles back to its input:



Exercise 4.1.8

What is the range of the identity function?

Exercise 4.1.9

What is the inverse of the identity function?

What can we say about the compositions of this special function, I, with another, g? Whether the other function comes before or after the identity function, the result is the same; this is our conclusion:

Theorem 4.1.10: Composition with Identity Function

A composition of any function with the identity function is that function, i.e.,

$$I \circ g = g, \quad g \circ I = g\,.$$

Proof.

Consider this diagram:

$$x \mapsto y = I(x) = x \implies x \mapsto g(y) = g(I(x)) = g(x)\,.$$

The output is the same as the input! Here is the other one:

$$y \mapsto x = g(y) \implies y \mapsto I(x) = I(g(y)) = g(y)\,.$$

Again, the output is the same as the input!

Numbers can be represented in a number of ways, but sometimes they are *identical*:

$$1 + 1 = 2\,.$$

Similarly, functions can be represented in a number of ways, but sometimes they are *identical*:

$$x + x = 2x\,.$$

Let's make it clear what we mean when we say that *two functions are equal* (or identical):

▶ Two functions are equal when each possible input produces the same output for either function.

Definition 4.1.11: equal functions

Suppose f and g are two functions:

$$f, g : X \to Y\,.$$

They are *equal*,

$$\boxed{f = g}$$

if their domains are equal (as sets) and we have the following:

$$f(x) = g(x) \quad \texttt{FOR EACH } x$$

It is illustrated in the flowchart below:

$$\begin{array}{rcccccccl} & & & \nearrow & x & \to & \boxed{f} & \to & y & \\ x & \to & x & & & & & & & \text{same!} \\ & & & \searrow & x & \to & \boxed{g} & \to & u & \end{array}$$

Example 4.1.12: equal functions

Consider these two functions:

$$f(x) = 2x + 4 \quad \text{and} \quad g(x) = 2(x + 2)\,.$$

No matter what x is, the outputs are the same. We conclude that they are equal: $f = g$. This is just as example of how we algebraically manipulate formulas.

What does it mean when we say that *two functions are not equal*? The opposite of equal: The outputs don't fully match. In other words, the answer is: A possible input produces two different outputs for the two functions.

Definition 4.1.13: not equal functions

Suppose f and g are two functions:

$$f, g : X \to Y\,.$$

They are *not equal*,

$$f \neq g$$

if their domains are unequal (as sets) or we have the following:

$$f(x) \neq g(x) \quad \texttt{FOR SOME } x$$

In the definition, we test each x: Do the values match? It is illustrated in the flowchart below:

$$\begin{array}{rcccccccl} & & & \nearrow & 3 & \to & \boxed{f} & \to & 7 & \\ 3 & \to & 3 & & & & & & & \text{different!} \\ & & & \searrow & 3 & \to & \boxed{g} & \to & 12 & \end{array}$$

As you can see, you only need to find a single value of x for which there is a mismatch.

This case also includes the situation when the two domains are unequal:

$$\begin{array}{rcccccccl} & & & \nearrow & 3 & \to & \boxed{f} & \to & 7 & \\ 3 & \to & 3 & & & & & & & \text{different!} \\ & & & \searrow & 3 & \to & \boxed{g} & \to & \text{“not in the domain!”} & \end{array}$$

Example 4.1.14: not equal functions

Consider these two functions:

$$f(x) = \frac{x^2}{x} \quad \text{and} \quad g(x) = x\,.$$

We conclude that they aren't equal: $f \neq g$. Why? Because f is undefined at $x = 0$, which is in the implied domain of g. Replacing f with g is an example of how *not* to do symbol manipulation!

Example 4.1.15: identities

This is a familiar identity:

$$(x+1)^2 = x^2 + 2x + 1\,.$$

A more complex two variable identity is

$$(x+y)^2 = x^2 + 2xy + y^2\,.$$

And so is

$$x^2 - y^2 = (x-y)(x+y)\,.$$

And so are all rules of exponents and a lot more presented in this chapter.

In the definition of the *inverse*, we take two round trips:

$$\begin{array}{lccccc} \text{start} & x & \to & \boxed{f} & \to & y \\ & \| & & & & \downarrow \\ \text{finish} & x' & \leftarrow & \boxed{f^{-1}} & \leftarrow & y \end{array} \qquad \begin{array}{ccccccc} x & \to & \boxed{f} & \to & y & & \text{finish} \\ \uparrow & & & & \| & & \\ x & \leftarrow & \boxed{f^{-1}} & \leftarrow & y' & & \text{start} \end{array}$$

Both times we arrive where we started – with the same final output. We interpret this definition in terms of the identity function, as follows.

Theorem 4.1.16: Inverse via Compositions

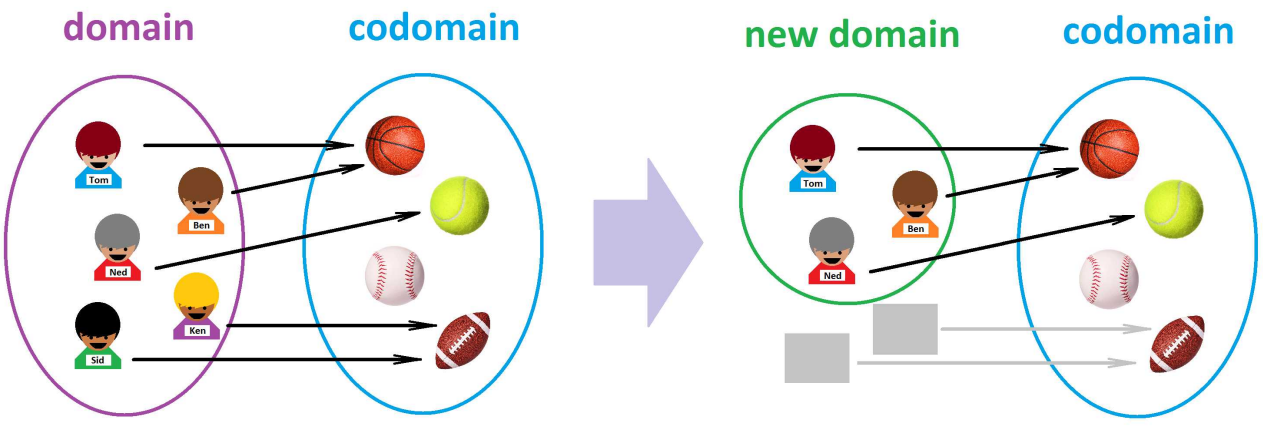
Two functions $f : X \to Y$ and $f^{-1} : Y \to X$ are inverses of each other if and only if their compositions produce the identity functions; i.e., these two conditions are satisfied:

$$f \circ f^{-1} = I \quad \textit{AND} \quad f^{-1} \circ f = I$$

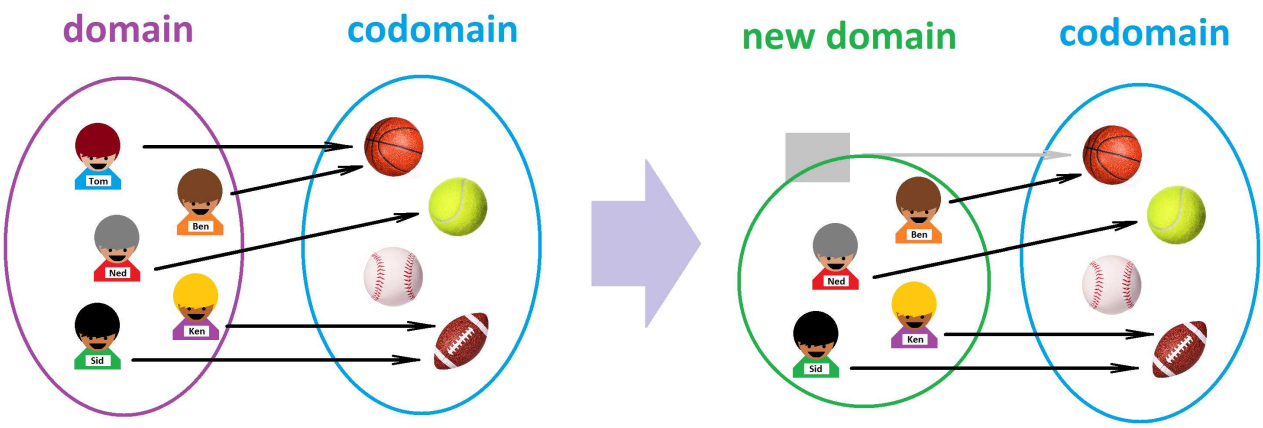
Warning!

There are *two* conditions here.

We have seen how "reducing" the domain of a function creates a new function:



Any subset of the domain can be chosen, but excluding Tom creates a function that is one-to-one:



Exercise 4.1.17

Change the codomain of the function to make it onto.

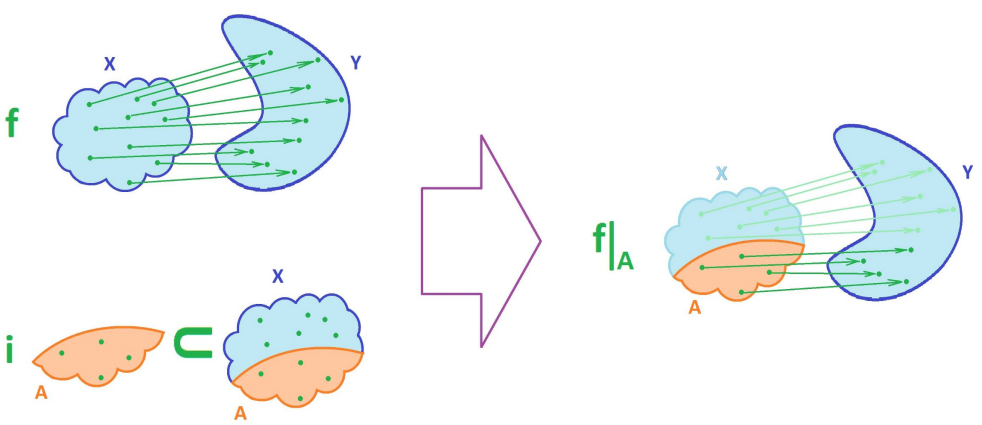
Definition 4.1.18: restriction of function

Suppose we have sets X and Y, a function $f : X \to Y$, and a subset A of X. Then the *restriction of f to A* is the function defined, and denoted, by the following:

$$f\Big|_A(x) = f(x) \quad \texttt{FOR EACH } x$$

in A.

The notation is reminiscent of the substitution notation. The construction is illustrated below:



Warning!

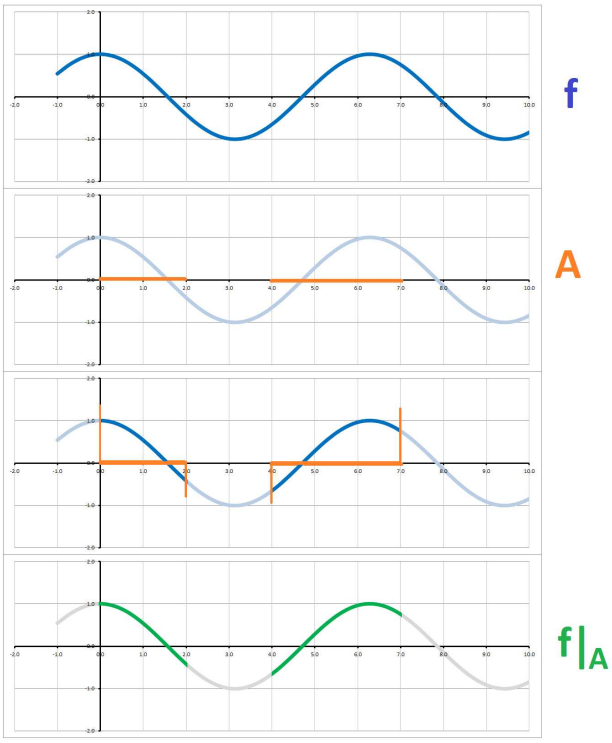
We have $f\Big|_A \neq f$, unless $A = X$.

Exercise 4.1.19

Explain how inclusions are restrictions of the identity functions.

Example 4.1.20: cosine restricted

In case of a numerical function, we can just erase any part of its graph, thus creating a new function. Here the cosine function is restricted to a set that consists of two intervals:

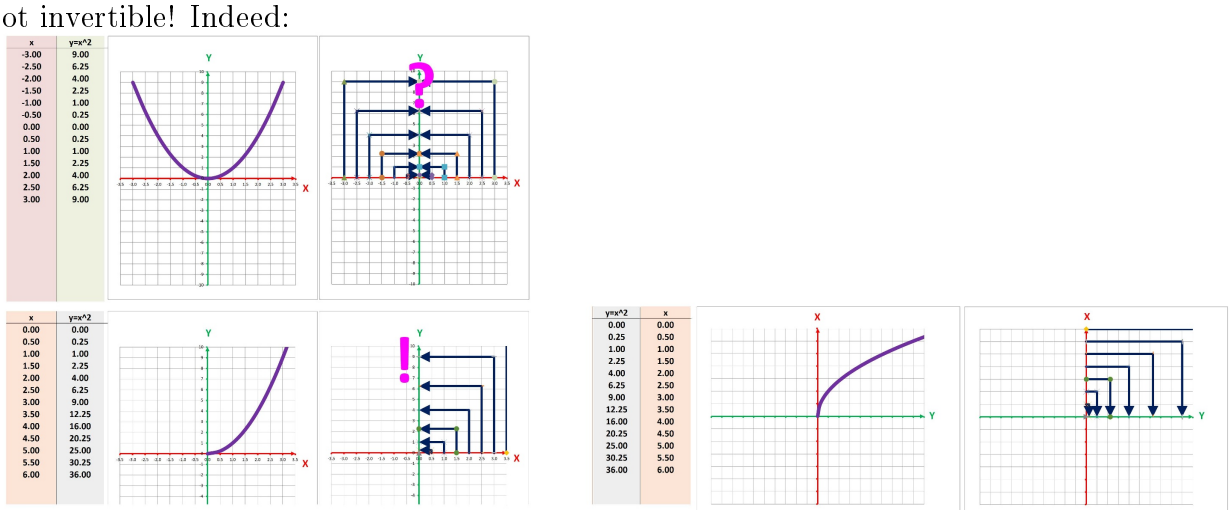


Example 4.1.21: restrictions improve functions

Restrictions allow us to create functions that are one-to-one from those that aren't. We then build "restricted" inverses of these functions. Recall how we did this in the last chapter. We have the square function:

$$f : \mathbf{R} \to \mathbf{R}, \quad f(x) = x^2 \,.$$

It's not invertible! Indeed:



Then, we just take one, of the two, branch of the graph (no matter which one):

$$g = f\Big|_{[0,+\infty)} \quad \text{or} \quad g = f\Big|_{(-\infty,0]} \,.$$

We have a new function,

$$g : [0,+\infty) \to [0,+\infty), \quad g(x) = x^2 \,,$$

and it *is* one-to-one! There is the inverse and it is the *square root* function:

$$g^{-1} : [0,+\infty) \to [0,+\infty), \quad g^{-1}(y) = \sqrt{y} \,.$$

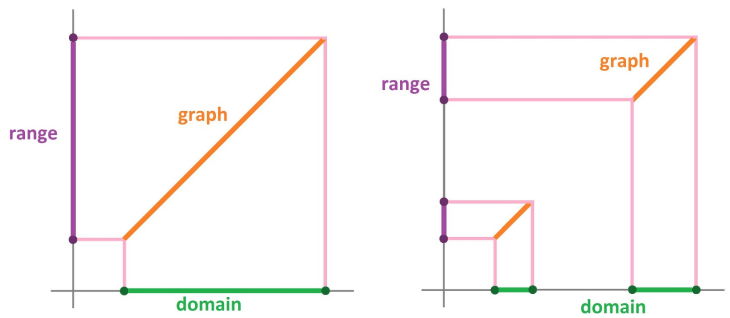
The other branch of x^2, and the other inverse, shouldn't be discarded though; it reappears every time we solve a quadratic equation:

$$x^2 = 2 \implies x = \pm\sqrt{2} \,.$$

Exercise 4.1.22

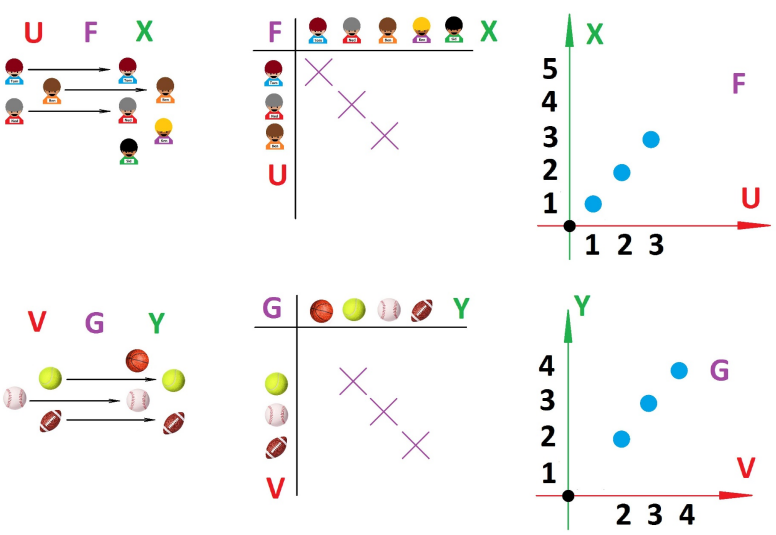
List all intervals A that satisfy: $f(x) = x^2$ restricted to A is one-to-one.

The graph of the *numerical* identity function is the diagonal line:



Of course, if the domain and, therefore, the codomain of this function are disconnected, then so is the graph.

The idea of such a function applies even if the two sets – domain and codomain – don't match. It suffices that the former is a subset of the latter. We modify the above example below:



We see U *included* in X, and V *included* in Y. This fact creates a function.

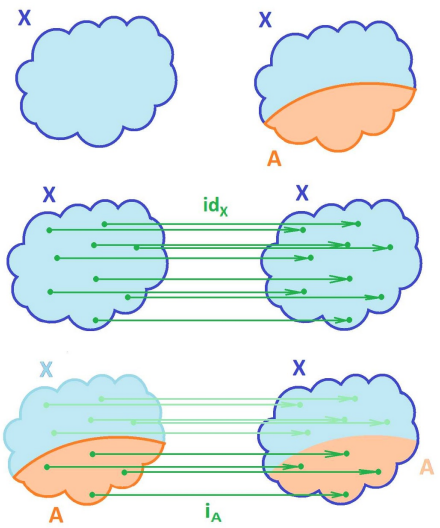
Definition 4.1.23: inclusion

Suppose we have a set X and a subset A of X. Then the *inclusion* $i : A \to X$ *of* A *into* X is the function defined by the following:

$$i(x) = x \quad \texttt{FOR EACH } x$$

in A.

Below, we make a copy of a set X with a subset A specified. We then construct the inclusion with identical, horizontal arrows:



Theorem 4.1.24: Restriction via Compositions

A restriction of a function is its composition with the appropriate inclusion; i.e., given a function $f : X \to Y$, a subset $A \subset X$, and its inclusion $i : A \to X$, the restriction of f to A is:

$$f\Big|_A = f \circ i$$

Exercise 4.1.25

Prove the theorem.

Exercise 4.1.26

A restriction of a constant function is ___ .

4.2. Monotonicity and the extreme values of functions

As we introduce new functions, we put them into broad categories. The main ones have been so far:

1. *bounded and unbounded*
2. *one-to-one* and *onto*

The next one is just as important.

When an object is moving up, we can also say that its altitude is *increasing*, and when it's falling down, the altitude is *decreasing*. We apply this language to all quantities and all functions.

When we say that a function increases, we mean that the output increases when the input does, and we say it decreases when the output decreases even though the input increases. In the former case, the graph *rises*; in the latter, it *drops* as we move from left to right:

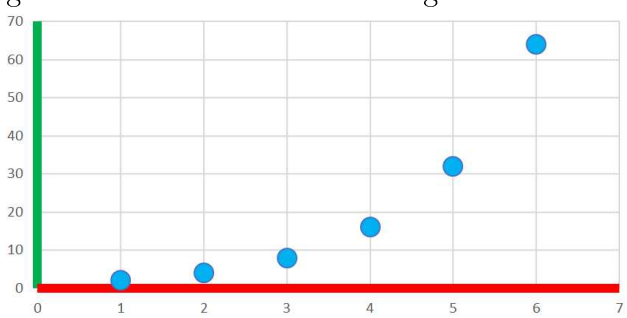


The graph tells the whole story, but we can also see a variety of behaviors within either category.

Example 4.2.1: monotone sequences

Sequences are functions, and the graphs above might be the graphs of sequences zoomed out! The idea of increasing and decreasing matches what we discussed in Chapter 1. As we know, an arithmetic progression is increasing when the increment is non-negative and decreasing when it's non-positive. A geometric progression is increasing when the increment is non-negative and the first element is positive:

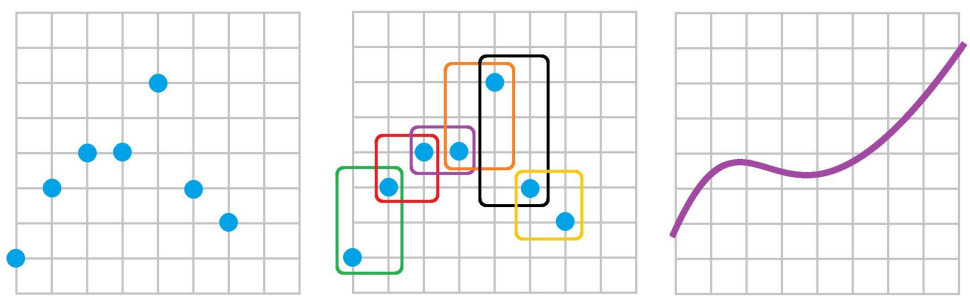
n	y
1	2.0
2	4.0
3	8.0
4	16.0
5	32.0
6	64.0



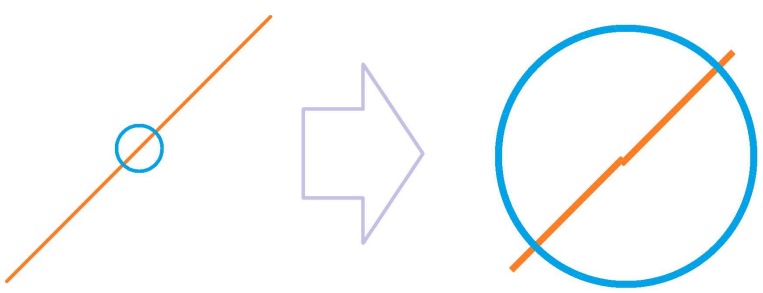
These conclusions are easy because we need only to compare two *consecutive* values of the input at a time:

$$a_n \leq a_{n+1} \quad \text{or} \quad a_n \geq a_{n+1}\,.$$

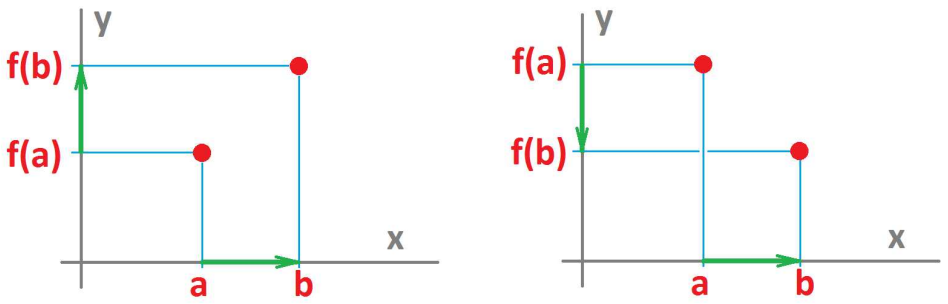
It's not as easy with functions with intervals as domains as there is no such thing as consecutive values:



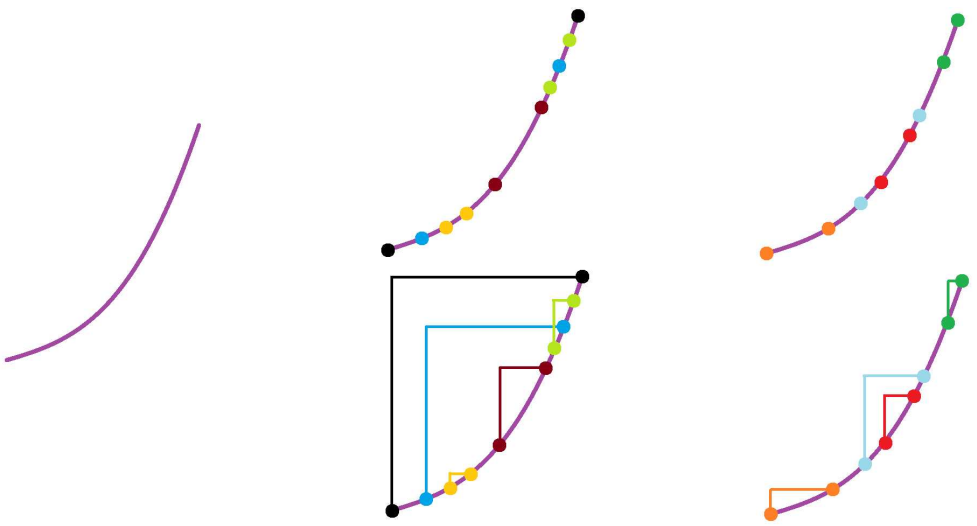
This verbal description is simple and its geometric meaning is clear. However, both are imprecise. What if we zoom in on the graph? Will every increasing function have its graph rising on a smaller scale? Maybe not:



We realize that we have just been describing the graph, which is nothing but a mere visualization of the function. The precise definition must rely on comparing *one pair of points at a time* ignoring the rest:



We have to face *infinitely many* pairs of points:



The following concepts are crucial.

Definition 4.2.2: increasing and decreasing functions

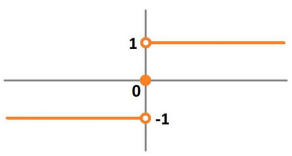
Suppose $y = f(x)$ is a function and an interval I lies within its domain. Then:

- The function f is called *increasing on the interval* I if, for every $a < b$ in I, we have:
$$f(a) \leq f(b)\,.$$
- The function f is called *decreasing on the interval* I if, for every pair $a < b$ in I, we have:
$$f(a) \geq f(b)\,.$$

Collectively, these functions are called *monotone*.

Example 4.2.3: sign function

Some step functions, such as the *sign function*, are increasing in spite of having a graph with horizontal parts:



In order to accommodate the possibility that these values cannot be equal to each other, we replace the non-strict inequality signs "$\leq$" and "$\geq$" above with strict: "$<$" and "$>$". The sign function is satisfies the former but not the latter.

Definition 4.2.4: strictly increasing and decreasing functions

Suppose $y = f(x)$ is a function and an interval I lies within its domain. Then:

- The function f is called *strictly increasing on the interval* I if, for every pair $a < b$ in I, we have:
$$f(a) < f(b)\,.$$
- The function f is called *strictly decreasing on the interval* I if, for every pair $a < b$ in I, we have:
$$f(a) > f(b)\,.$$

Collectively, these functions are called *strictly monotone*.

Warning!

A constant function is both increasing and decreasing but it's not strictly increasing or decreasing.

Exercise 4.2.5

Insert the correct symbol:

$$f \text{ is increasing } \_\_\_ \ f \text{ is strictly increasing }.$$

We will use these arrows to indicate the type of the function's monotonicity:

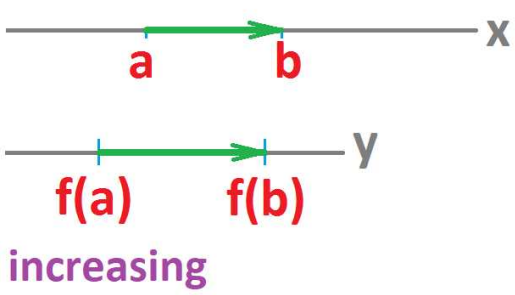
Monotonicity

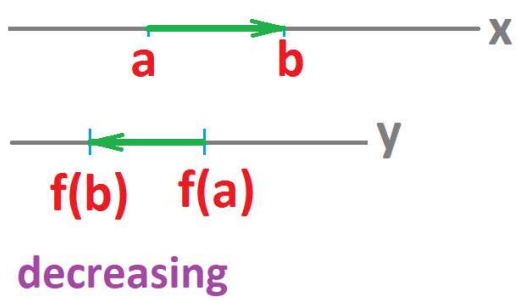
- " $\nearrow$ " increasing
- " $\searrow$ " decreasing

The definition takes the following form:

$\nearrow$	$a \leq b$	$\Longrightarrow$	$f(a) \leq f(b)$
$\searrow$	$a \leq b$	$\Longrightarrow$	$f(a) \geq f(b)$

What does this concept means when we look at functions as *transformations*? Let's place the x- and the y-axis side by side:





This is our conclusion:

▶ An increasing function doesn't flip any parts of the interval and a decreasing flips all.

Exercise 4.2.6

What does a vertical shift do to the monotonicity of a function? A horizontal shift? The flips? The stretches?

How do we verify these conditions? We can utilize what we know about *the algebra of inequalities* (more on those in Chapter 5).

Example 4.2.7: monotonicity and inequalities

First, we can add any number to both sides of an inequality:

$$a < b \implies a + 4 < b + 4 \,.$$

Therefore, the function $f(x) = x + 4$ is increasing.

We can multiply both sides of an inequality by a positive number:

$$a < b \implies 3a < 3b \,.$$

It follows that the function $f(x) = 3x$ is increasing.

Third, if we multiply both sides of an inequality by a negative number, we have to reverse the sign:

$$a < b \implies (-2)a > (-2)b \,.$$

It follows that the function $f(x) = -2x$ is decreasing.

Below is the rule we use for doing algebra with inequalities:

Rule of Signs for Multiplication

$$\begin{array}{rcl} (+) \cdot (+) & = & (+) \\ (+) \cdot (-) & = & (-) \\ (-) \cdot (+) & = & (-) \\ (-) \cdot (-) & = & (+) \end{array}$$

Example 4.2.8: linear monotonicity

This is how we can solve this problem one function at a time, from scratch. Let

$$f(x) = 3x - 7 \,.$$

Compare:

$$\begin{array}{rrcl} f(x_1) = & 3x_1 - 7 & \overset{?}{<} & f(x_2) = 3x_2 - 7 \\ \iff & 3x_1 & \overset{?}{<} & 3x_2 \\ \iff & x_1 & < & x_2 \,. \end{array}$$

The computation proves that this function is increasing.

Putting these ideas together, we acquire the following:

Theorem 4.2.9: Monotonicity of Linear Functions

A linear polynomial is

1. *increasing when its slope is positive, and*
2. *decreasing when its slope is negative.*

In other words:

1. $f(x) = mx + b \ \nearrow \iff m > 0 \,.$
2. $f(x) = mx + b \ \searrow \iff m < 0 \,.$

Exercise 4.2.10

Prove the theorem.

Exercise 4.2.11

What is the missing part?

Things get harder for the quadratic, cubic, ... powers as they lead to quadratic, cubic, ... inequalities.

Example 4.2.12: monotonicity of x^2

Let's consider

$$f(x) = x^2 \,.$$

First, we can multiply two inequalities, when they are aligned and their signs are positive:

$$\begin{array}{rcccl} 0 & < & a & < & b \\ 0 & < & a & < & b \\ \hline & & a^2 & < & b^2 \end{array}$$

Therefore, the function $f(x) = x^2$ is increasing on the interval $[0, +\infty)$. Second, if we multiply two inequalities when their signs are negative, we have to reverse the sign:

$$\begin{array}{rcccl} a & < & b & < & 0 \\ a & < & b & < & 0 \\ \hline a^2 & > & b^2 & & \end{array}$$

Therefore, the function $f(x) = x^2$ is decreasing on the interval $(-\infty, 0]$.

Exercise 4.2.13

How does one address the monotonicity of quadratic functions in general?

Example 4.2.14: monotonicity of cubic function

Now, we let

$$f(x) = x^3 \,,$$

and follow a similar procedure starting with unknown a and b with $a < b$. We can multiply *three* identical inequalities – positive or negative – and preserve the sign:

$$\begin{array}{rcl} a & < & b \\ a & < & b \\ a & < & b \\ \hline a^3 & < & b^3 \end{array}$$

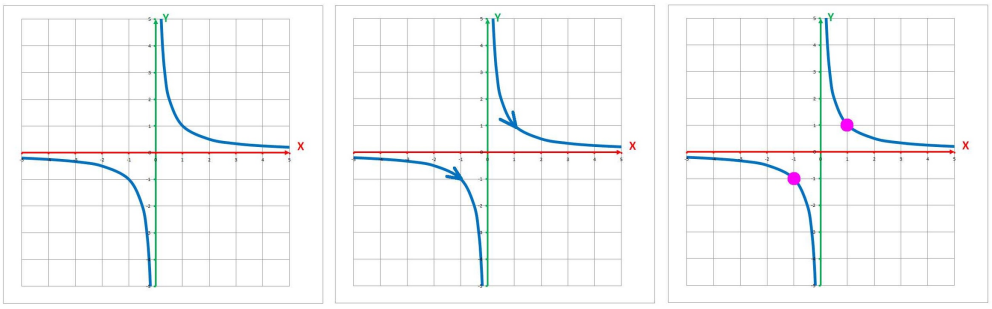
Therefore, the function $f(x) = x^3$ is increasing on the interval $(-\infty, +\infty)$.

Example 4.2.15: monotonicity of reciprocal function

The reciprocal is a bit tricky:

$$y = \frac{1}{x} \,.$$

Indeed, while both branches go *down*, crossing the y-axis brings the values *up*!



The challenge stems from the fact that its domain has a *hole*. Therefore, the definition can only apply to an interval I that lies entirely to the left of 0 or to the right, such as

$$I = (-\infty, 0) \quad \text{or} \quad I = (0, +\infty)\,.$$

On these two intervals – seen separately – the function is *decreasing*.

Exercise 4.2.16

Prove the last statement algebraically.

Exercise 4.2.17

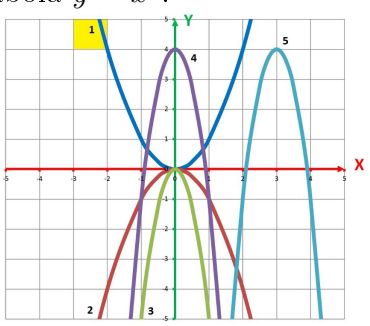
How do the basic transformations of the plane affect the monotonicity of a function?

Example 4.2.18: monotonicity of quadratic polynomials

Things become much more complex if we need to analyze a general quadratic function:

$$f(x) = ax^2 + bx + c,\ a \neq 0\,.$$

We recall that all quadratic functions are represented by *parabolas* (Chapter 3). Each of them is a result of a transformation of *the* parabola $y = x^2$.



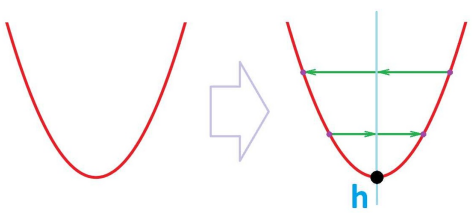
From the above example, we know:

▶ The function $y = x^2$ is strictly decreasing on $(-\infty, 0)$ and strictly increasing on $(0, +\infty)$.

We know also that we need only these four transformations:

- The vertical stretch doesn't make the function change its monotonicity on any interval.
- The vertical shift doesn't make the function change its monotonicity on any interval.
- The vertical flip makes the function change its monotonicity to the opposite.
- The horizontal shift shifts the intervals of monotonicity of the function.

The last transformation determines the location of the *vertex* of the parabola, the point that separates increasing from decreasing behavior:



Its x-coordinate is

$$h = -\frac{b}{2a}\,.$$

Then, we have a classification of all quadratic functions:

- If $a > 0$, then f is strictly decreasing on $(-\infty, h)$ and strictly increasing on $(h, +\infty)$.
- If $a < 0$, then f is strictly increasing on $(-\infty, h)$ and strictly decreasing on $(h, +\infty)$.

Thinking geometrically, the graph of a strictly increasing function can't come back and cross a horizontal line for the second time:



It, therefore, passes the *Horizontal Line Test*. The result below follows.

Theorem 4.2.19: Monotone vs. One-to-one

All strictly monotone functions are one-to-one.

Proof.

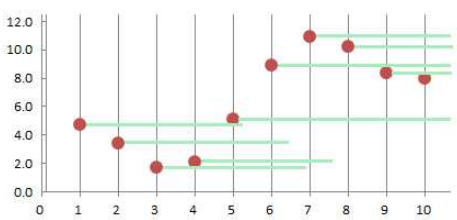
Algebraically, we start with the definition and conclusion follows:

$$f \nearrow \text{ when: } \underbrace{x_1 < x_2}_{\text{two different inputs}} \implies f(x_1) < f(x_2) \implies \underbrace{f(x_1) \neq f(x_2)}_{\text{two different outputs}}\,.$$

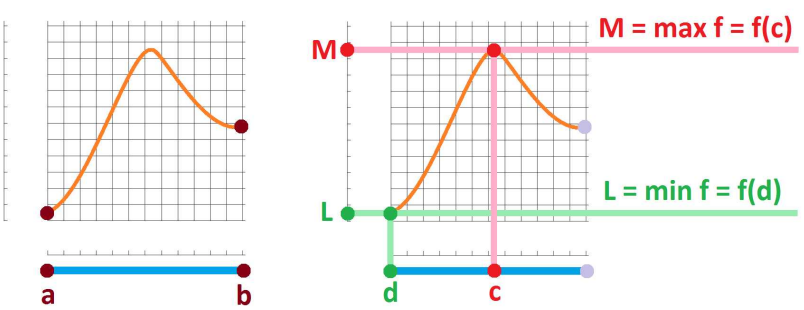
Hence f is one-to-one.

Exercise 4.2.20

Some functions, such as sequences, might *skip values*. What does the graph below tell us about the theorem?



Recall from Chapter 2 that *bounded functions* are defined as one that have bounded images (ranges). Optimizing a function, therefore, means finding one or both of the most important bounds of the image: *maximum and minimum*.



In the spirit of this section, we can give another definition of this concept.

Definition 4.2.21: maximum and minimum of function

Suppose $y = f(x)$ is a function and an interval I lies within its domain. Then:

- A point $x = d$ in I is called a *global maximum point* of f on I if

$$f(d) \geq f(x) \quad \texttt{FOR EACH } x$$

in I.

- A point $x = c$ in I is called a *global minimum point* of f on I if

$$f(c) \leq f(x) \quad \texttt{FOR EACH } x$$

in I.

Collectively, they are all called *global extreme points*. Furthermore, $M = f(d)$ and $m = f(c)$ are called *global maximum and minimum values* (and global extreme points) of f on I, respectively.

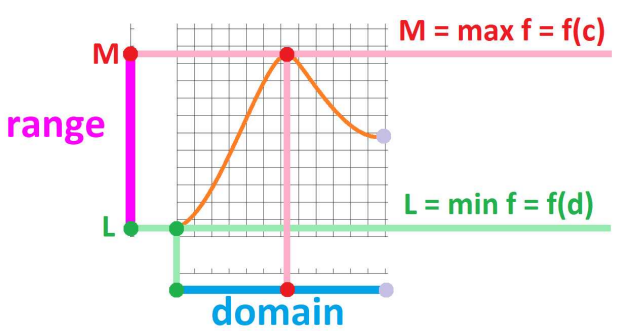
Warning!

Some sources call these *absolute* maximum and minimum points/values.

Warning!

"Maximum", "minimum", and "extremum" may refer to the x-value, or the y-value, or the (x, y)-point on the graph.

We now use the same words for *functions* as we used for sets. The reason is that the extrema of a function are those of its range:



In fact, above we have:

$$\text{Range} \quad = [L, M]\,.$$

Example 4.2.22: extrema of quadratic polynomials

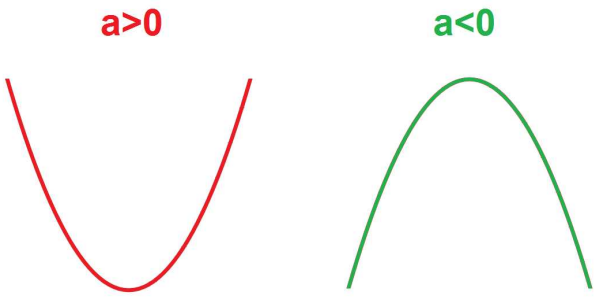
The graph of a quadratic function,

$$f(x) = ax^2 + bx + c, \ a \neq 0\,,$$

is a parabola. Therefore, the *vertex* of this parabola is the maximum or minimum of this function. In either case, its x-coordinate is

$$h = -\frac{b}{2a}\,.$$

In the meantime, the sign of a tells the two cases apart:



We have also demonstrated – beyond a conjecture – that the largest possible enclosure the farmer (Chapter 2) can build from 100 yards of fencing will have the area of 625 square yards.

Warning!

For a constant function, every point is both a maximum and a minimum point.

Exercise 4.2.23

(a) How many maximum and minimum values can a function possibly have? (a) How many maximum and minimum points can a function possibly have?

Exercise 4.2.24

Identify extreme points of the functions discussed in this section.

Exercise 4.2.25

Define "strict extreme points".

In summary, we are proving one of these inequalities:

	x_1	$<$	x_2
$f \nearrow$	$f(x_1)$	$\leq$	$f(x_2)$
$f \searrow$	$f(x_1)$	$\geq$	$f(x_2)$

and

			c		x
c	is	min	$f(c)$	$\leq$	$f(x)$
c	is	max	$f(c)$	$\geq$	$f(x)$

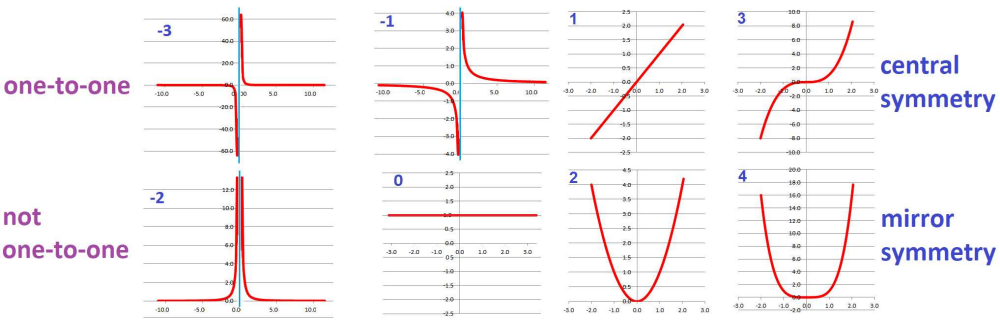
4.3. Functions with symmetries

The graphs of some of the functions that we have discussed so far have discernible features. As we introduce new functions, we continue to put them into broad categories according to those features. The main ones we have seen so far (possibly limited to a subset of the domain) are the following:

1. bounded and unbounded
2. one-to-one and onto
3. increasing and decreasing

The next one is just as important.

Every one of the eight power functions below has a *symmetry*, and which one it is depends on the degree:



If the pieces of paper with the graphs in the second row (even powers) are folded along the y-axis, the left and the right branches of each graph will *merge*. This won't happen with the graphs in the first row (odd powers) even though the two branches seem identical... They do merge if we rotate them 180 degrees.

A subset of the real line or the xy-plane has a symmetry if we can turn it into itself via some *motion* of the line or the plane, respectively. Of course, motions are just a special kind of *transformation*.

Just as in Chapter 3, we start with the transformations of the *line*.

We will consider only these two:

1. the shift, $F(x) = x + b$, the translation by b, and
2. the flip, $F(x) = -x$, reflection about 0.

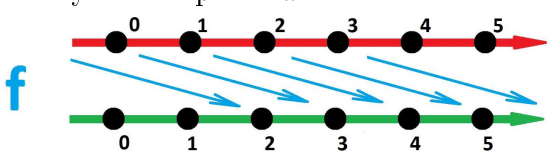
We exclude the stretch because it's not a rigid motion.

Example 4.3.1: translation symmetry

We know that if n and m are integers, then so is $n + m$. Therefore, the set of integers $\mathbf{Z}$ is symmetric with respect to an *integer* shift. Furthermore, the set of all integer multiples of a fixed real number, say a,

$$A = \{ka : \ k = \ldots -2, -1, 0, 1, 2, 3, \ldots\},$$

translates into itself when shifted by a multiple of a:



Of course, the set of reals and the set of rational numbers have this symmetry.

Exercise 4.3.2

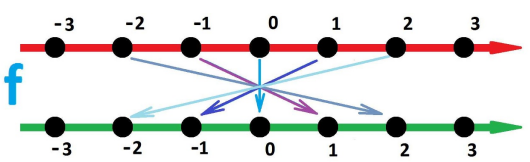
Does the set of irrational numbers have this symmetry?

Exercise 4.3.3

Prove that such a set would have to be infinite.

Example 4.3.4: reflection symmetry

We know that if n is an integer, then so is $-n$. Therefore, the set of integers $\mathbf{Z}$ is symmetric with respect to the *flip* shift:



Other examples of sets that are symmetric relative to the flip are these:

- $\{0\}$
- $\{-r, r\}$

- $(-r, r)$
- $[-r, r]$

Exercise 4.3.5

For the two symmetries above, give more examples of a subset of $\mathbf{R}$ that do or do not have it.

Now the transformations of the *plane*.

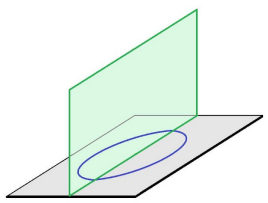
We will consider only these four:

1. the vertical shift, $y \to y + b$, the translation by b,
2. the horizontal shift, $x \to x + b$, the translation by b,
3. the vertical flip, $y \to -y$, reflection about the x-axis,
4. the horizontal flip, $x \to -x$, reflection about the y-axis.

We exclude the stretches because they are not rigid motions.

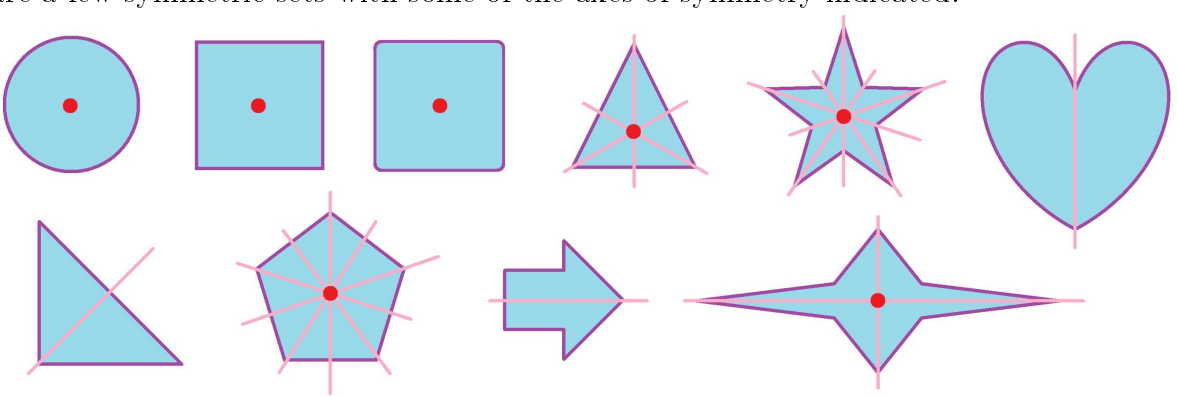
Example 4.3.6: axes of symmetry

A circle is drawn on a piece of paper and then a piece of glass is put vertically through its center. But, is this a piece of transparent glass or a mirror?



The symmetry is such that we see a circle no matter what. The line on which the mirror stands is called the "axis of symmetry". Of course, we can place the mirror at any angle to have this effect as long it goes through the center of the circle.

These are a few symmetric sets with some of the axes of symmetry indicated:



As we can see, some sets have more than one axis of symmetry. Meanwhile, the red dots indicate the centers of central symmetry.

Exercise 4.3.7

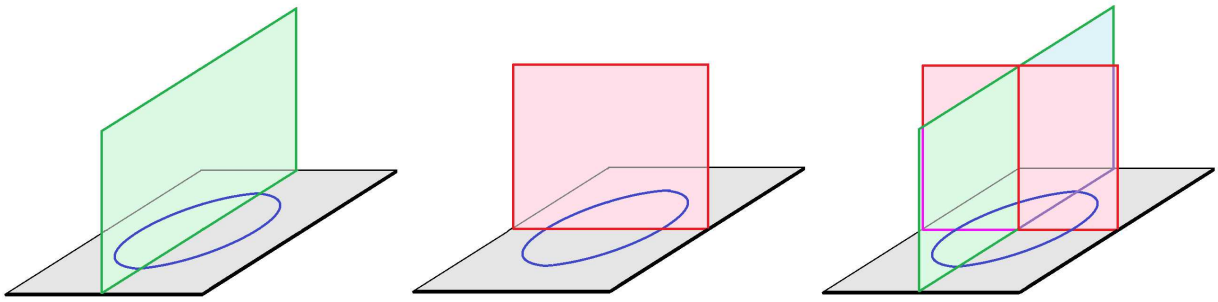
Provide all axes of symmetry for the circle and the squares.

Exercise 4.3.8

Does a point of intersection of two axes of symmetry always indicate a central symmetry?

Example 4.3.9: axes of symmetry

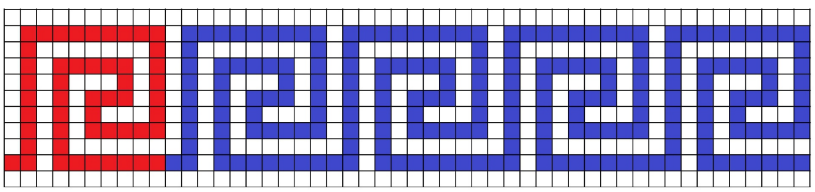
This is what happens if we put a second mirror on top of a circle drawn on a piece of paper:



We see a reflection in a reflection as if this isn't a mirror but a transparent glass. In addition to the two mirror symmetries, we can see the central symmetry.

Example 4.3.10: translation symmetry

This pattern below repeats itself under a certain horizontal shift:



Exercise 4.3.11

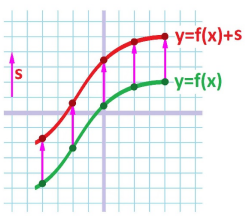
What other symmetries does it have?

Exercise 4.3.12

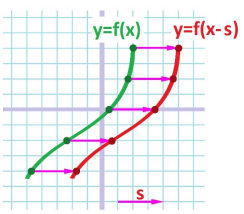
Explain how to make a mirror symmetric object from any object. What about centrally symmetric? Translation symmetric?

In this section, we will limit ourselves to the transformations of the plane that result from transformations of y-axis and the x-axis, or both, just as in Chapter 3. We will also concentrate on motions: shifts and flips. Furthermore, our interest here is the *graphs* of functions!

First, a vertical shift. The graph of no function will ever be transformed into itself with this shift as it would violate the *Vertical Line Test*:



The *horizontal shift* looks more promising:



It is given by:

$$(x, y) \xrightarrow{\text{right } s} (x + s, y)$$

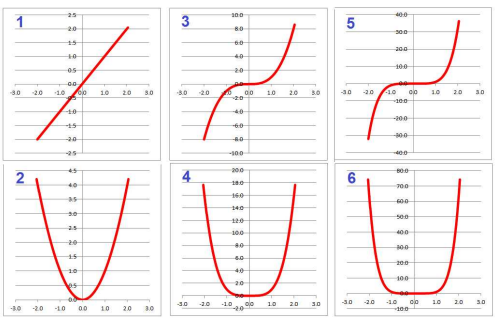
Furthermore, if the graph of G is the graph of f shifted k units to the right, then

$$G(x) = f(x - k)\,.$$

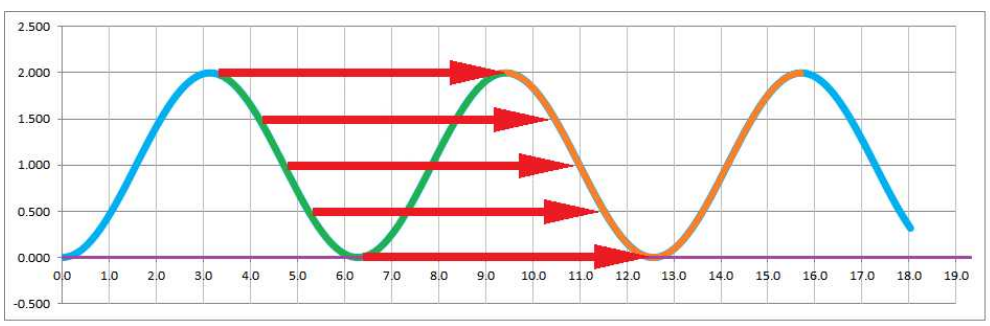
The new function must be the original one, f, for there to be symmetry.

Example 4.3.13: repetitive graphs

Since the identity must hold for every number x, the domain is infinite. Of course, we have seen many such functions. However, they all increase or decrease and, therefore, were not repetitive, such as the power functions:



In contrast, we can see how a piece of the graph shown below of the length of about 6 is repeated on and on:



The following concept will be commonly used.

Definition 4.3.14: periodic function

A function f is called *periodic* if there is such a $T > 0$ that

$$f(x + T) = f(x) \quad \texttt{FOR EACH } x$$

The smallest such T is called the *period* of f.

Such a function is repetitive: it goes over the same value in the same manner. That is why the main reason to study periodic functions is to represent idealized *repetitive processes*, such as these:

- the temperature during the year in a particular location
- the daylight over the year, the time of the sunset and the sunrise
- the oscillation of a string
- sound waves, etc.

Exercise 4.3.15

Describe such a function as a transformation.

Example 4.3.16: constant is periodic

Constant functions are – trivially – periodic, for any and every T:

$$f(x) = c = f(x+T)\,.$$

Exercise 4.3.17

What is its period?

Examples beyond the trivial are harder to come by, for now. For a function $y = f(x)$ given by a formula, we need to start with $f(x+T)$ and work our way to $f(x)$, identically. For example, no linear polynomial $f(x) = mx + b,\ m \neq 0$, is periodic:

$$f(x+T) = m(x+T) + b = mx + mT + b = f(x) + mT \neq f(x)\,,$$

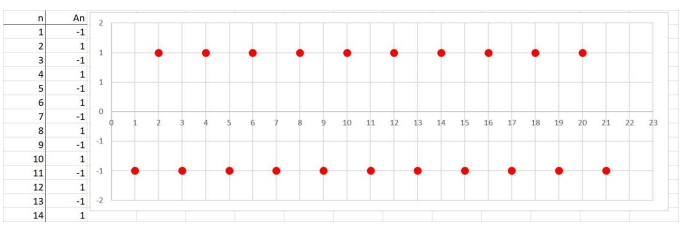
no matter the choice of T.

Example 4.3.18: alternating sequence is periodic

We can, however, pick the *alternating sequence*:

$$a_n = (-1)^n\,,$$

defined over all integers:



n	An
1	-1
2	1
3	-1
4	1
5	-1
6	1
7	-1
8	1
9	-1
10	1
11	-1
12	1
13	-1
14	1

It is periodic with period $T = 2$.

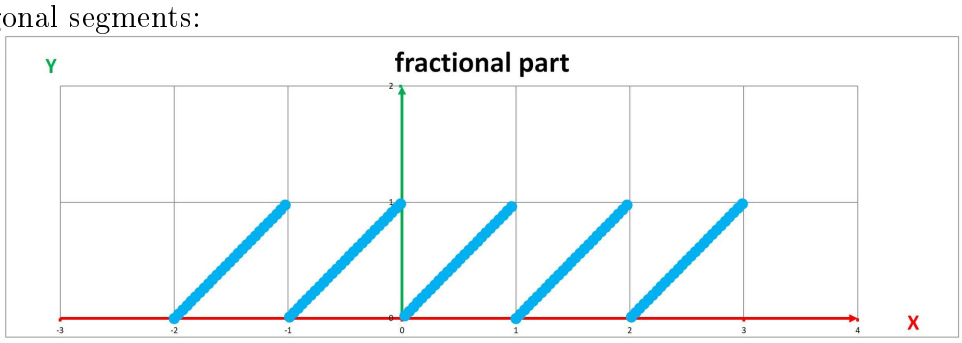
Example 4.3.19: fractional part

Another such function – with the domain all reals – is the so-called "fractional part":

$$\{x\} = x - [x]\,.$$

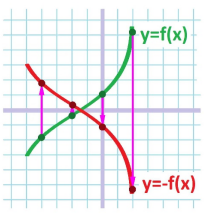
Its graph consists of diagonal segments:

x	y=[x]	y={x}
-2.00	-2	0.00
-1.96	-2	0.04
-1.93	-2	0.08
-1.89	-2	0.11
-1.85	-2	0.15
-1.81	-2	0.19
-1.78	-2	0.23
-1.74	-2	0.26
-1.70	-2	0.30
-1.66	-2	0.34
-1.63	-2	0.38
-1.59	-2	0.41

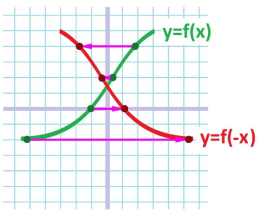


Its period is 1. More examples – the trigonometric functions – will appear later in the chapter.

What does the vertical flip do to graphs? With the exception of $f(x) = 0$, the graph of no function will ever be transformed into itself with this flip as it would violate the *Vertical Line Test*:



The *horizontal flip* looks more promising:



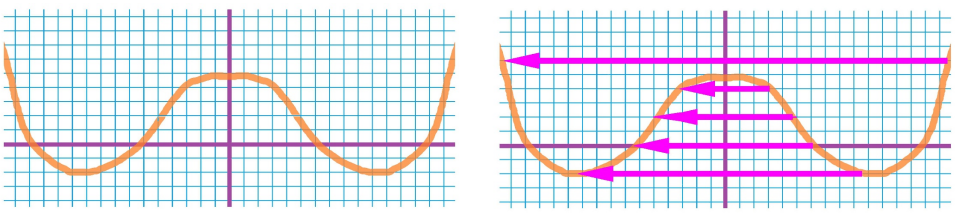
It preserves y and flips the sign of x:

$$(x, y) \xrightarrow{\text{horizontal flip}} (-x, y)$$

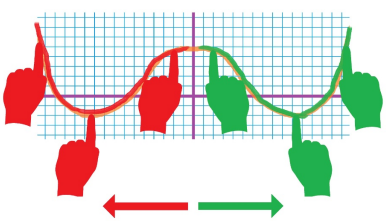
So, if the graph of $y = G(x)$ is the graph of $y = f(x)$ flipped horizontally, then

$$G(x) = f(-x)\,.$$

The new function must be the original one, f, for there to be symmetry:



We can also take advantage of the fact that our bodies have a mirror symmetry; one can trace the two halves of the graph with his two hands moving away from the center:



We have an identity to be verified. The following will be routinely used.

Definition 4.3.20: even function

A function f is called *even* if it satisfies:

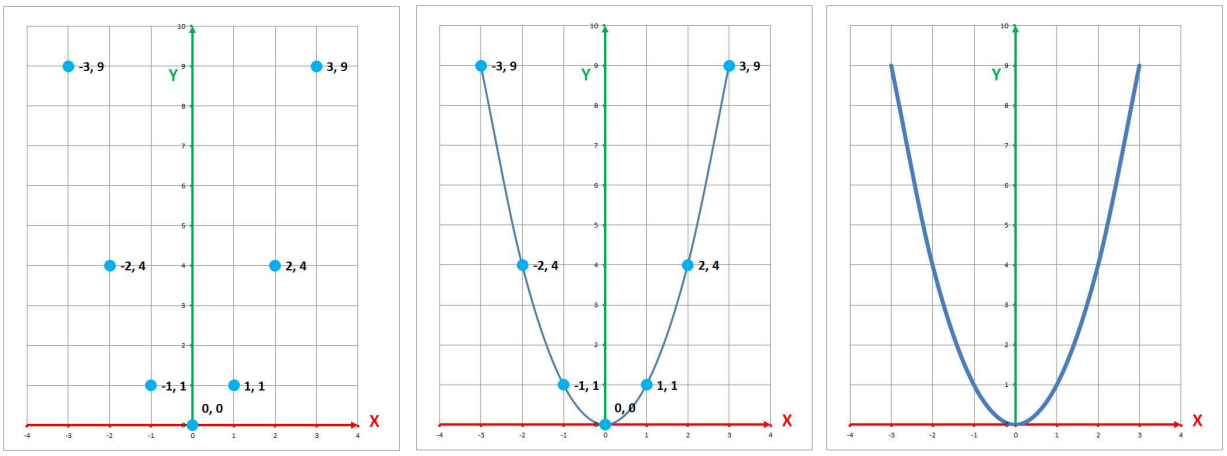
$$f(-x) = f(x) \quad \texttt{FOR EACH } x$$

in its domain.

The definition states that these two functions are the same!

Example 4.3.21: symmetry of squaring function

The graph of $f(x) = x^2$ seems to exhibit the horizontal mirror symmetry:



However, let's examine the *algebra*. For a function $y = f(x)$ given by a formula, we need to start with $f(-x)$ and work our way to $f(x)$, for all x's at the same time. The fact that the squaring function is even is proven algebraically as follows:

$$f(-x) = (-x)^2 = (-x) \cdot (-x) = (-1)x \cdot (-1)x = (-1)(-1)x^2 = x^2 = f(x)\,. \quad \text{Yes!}$$

Exercise 4.3.22

Prove the same for other even powers.

Exercise 4.3.23

Prove the same for the absolute value function

$$|-x| = x\,.$$

Exercise 4.3.24

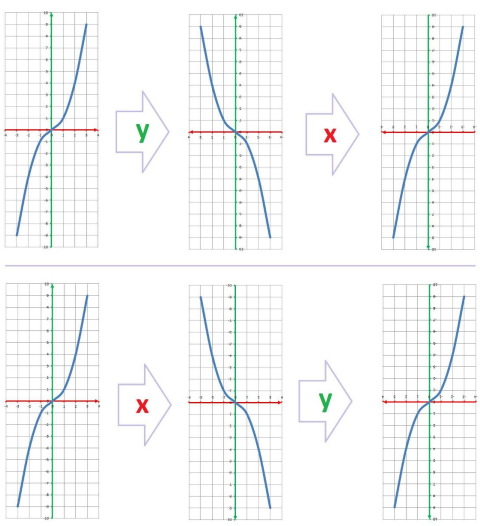
Describe an even function as a transformation.

Example 4.3.25: non-symmetry of cubic power

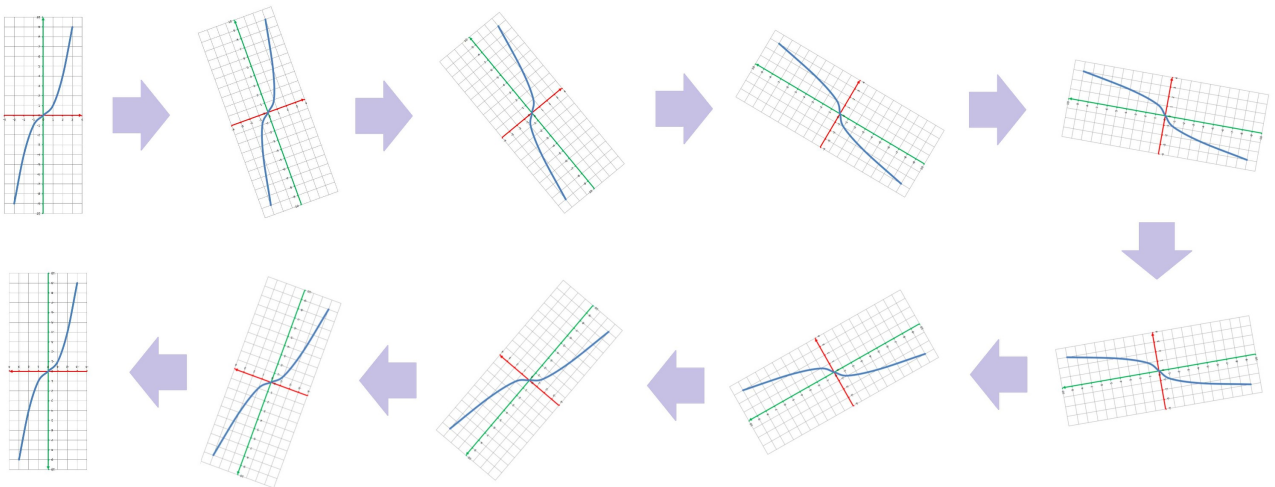
In the above example, the negative sign disappears. This won't happen with the cubic power. Indeed, we have a mismatch for every $x \neq 0$:

$$f(-x) = (-x)^3 = (-x) \cdot (-x) \cdot (-x) = (-1)^3 x^3 = -x^3 \neq x^3 = f(x)\,. \quad \text{No!}$$

Next, the *central symmetry* can be expressed as *two consecutive mirror symmetries*. Indeed, we just fold the plane in half along the y-axis and then along the x-axis to make one branch land on top of the other:



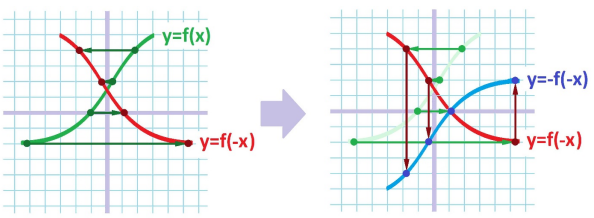
The two of them combined don't give a mirror image anymore; it's understood also as a *rotation through* 180 *degrees* around the origin:



The combination of vertical and horizontal flips "flips" the signs of both x and y:

$$(x, y) \xrightarrow{\text{central symmetry}} (-x, -y)$$

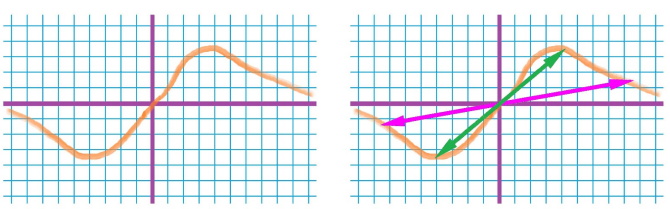
Below, we go from green to red and then from red to blue:



So, if the graph of $y = G(x)$ is the graph of $y = f(x)$ flipped horizontally and vertically, then

$$G(x) = -f(-x)\,.$$

The new function must be the original one, f, for there to be symmetry:



We have an identity to be verified. The following will be routinely used.

Definition 4.3.26: odd function

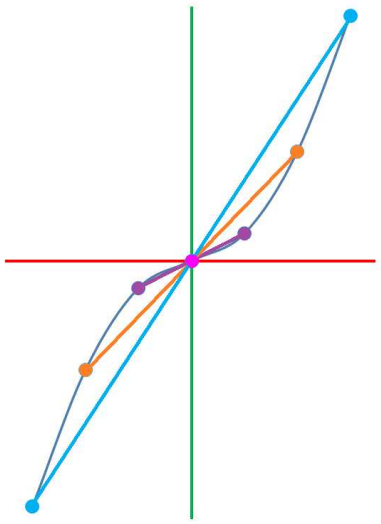
A function f is called *odd* if it satisfies:

$$f(-x) = -f(x) \quad \texttt{FOR EACH } x$$

in its domain.

Example 4.3.27: symmetry of cubic function

The graph of $f(x) = x^3$ seems to exhibit central symmetry:



However, let's examine the *algebra*. Once again, if we have a formula for a function $y = f(x)$, we need to start with $-f(-x)$ and work our way to $f(x)$, identically. The cubic power is odd:

$$-f(-x) = -(-x)^3 = -(-x) \cdot (-x) \cdot (-x) = -(-1)^3 x^3 = x^3 = f(x)\,. \quad \text{Yes!}$$

Exercise 4.3.28

Prove the same for other *odd* powers.

Exercise 4.3.29

Prove that the sign function is also odd:

$$-\operatorname{sign}(-x) = \operatorname{sign}(x)\,.$$

Exercise 4.3.30

Describe an odd function as a transformation.

Example 4.3.31: non-symmetry of square function

In the above proof, the negative sign is preserved, by design. This can't happen with the square power. Indeed, we have a mismatch for every $x \neq 0$:

$$-f(-x) = -(-x)^2 = -(-x) \cdot (-x) = -(-1)x \cdot (-1)x = -(-1)(-1)x^2 = -x^2 \neq x^2 = f(x). \quad \text{No!}$$

So, an even number of of (-1)'s produce a 1:

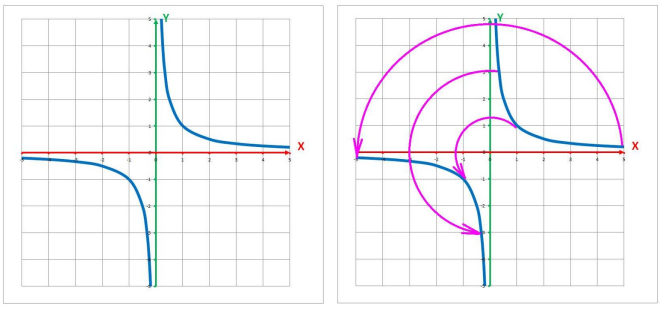
$$(-1)(-1) = 1\,.$$

An odd number of of (-1)'s produce a -1:

$$(-1)(-1)(-1) = -1\,.$$

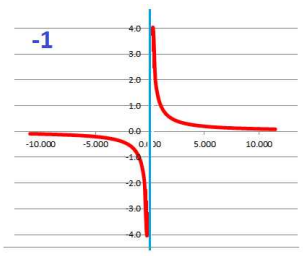
Example 4.3.32: symmetry and non-symmetry of reciprocal

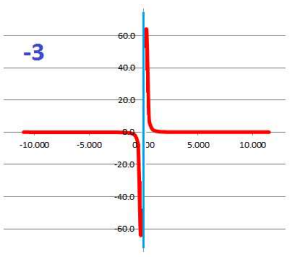
The function $1/x$ seems to have this symmetry:



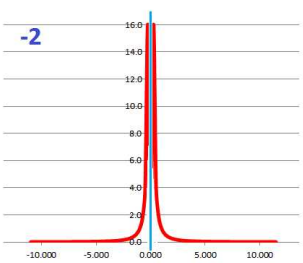
We can see it in the case of the rest of the reciprocals of the powers, which also confirms the reason for the names:

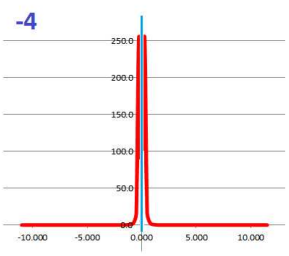
odd





even





Warning!

A randomly chosen, or a typical, function is neither even nor odd.

Exercise 4.3.33

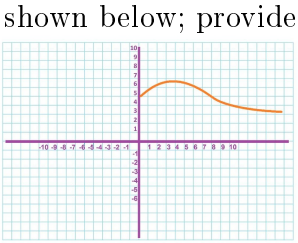
Demonstrate: (a) A vertical stretch will not preserve the graph of a function except for ___ . (b) A horizontal stretch will not preserve the graph of a function except for ___ .

Exercise 4.3.34

Is the composition of two odd/even functions odd/even?

Exercise 4.3.35

Half of the graph of an even function is shown below; provide the other half:



Hint: Try to trace with your left hand what your right hand is tracing.

Exercise 4.3.36

Half of the graph of an odd function is shown above; provide the other half.

Exercise 4.3.37

Plot the graph of a function that is both odd and even.

Exercise 4.3.38

Is the inverse of an odd/even function odd/even?

Exercise 4.3.39

What is the relation, if any, between these two pairs of properties: (a) even and one-to-one, (b) odd and one-to-one?

Exercise 4.3.40

What transformations of the plane preserve the "evenness" and "oddness" of functions?

Our list of the most fundamental classes of functions has grown to the following:

1. bounded and unbounded
2. one-to-one and onto
3. increasing and decreasing
4. odd and even
5. periodic

The functions introduced below (and their restrictions) will be matched against this list.

4.4. Quadratic polynomials

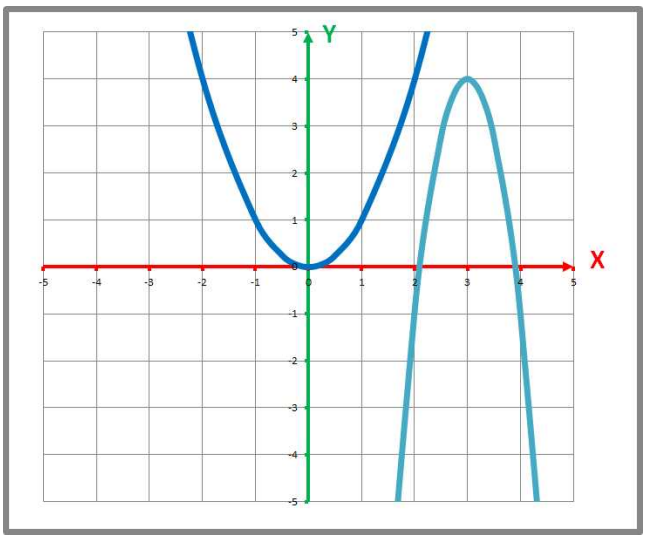
In Chapter 2, we thoroughly studied the "square function" $f(x) = x^2$. We now take the idea of squaring the input to the next level and combine it with other algebraic operations. A *quadratic polynomial* is a function given by

$$f(x) = ax^2 + bx + c, \ a \neq 0$$

We say that it is presented here in the *standard form*. The numbers a, b, c are the parameters that determine a lot of how the function behaves. We require the first coefficient a to be non-zero to separate the quadratic polynomials from the linear polynomials.

As we saw in Chapter 3, the graph of every quadratic polynomial is made from the "original parabola" of $y = x^2$ via shifting, flipping, and stretching. Some things about this polynomial will be easily deduced from those about the original that we learned in Chapter 2 but others will require some work.

The domain, the possible input values, remains the same, all reals, $(-\infty, +\infty)$. The output values can now be both positive and negative:



Just as with the original parabola, we notice that the graph above shows the lack of the undesirable features: gaps and breaks, corners and cusps...

Different inputs can produce, again, same outputs; quadratic polynomials aren't one-to-one! In fact, a pattern starts to emerge if we put the values of, say,

$$f(x) = x^2 - 2x + 2\,,$$

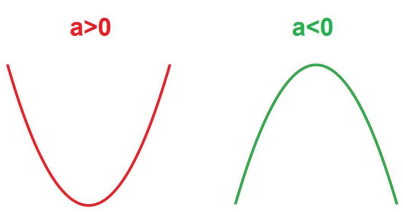
in a list:

$$\text{—— different inputs } \left\{ \begin{matrix} -2\ 10 \\ \left\{ \begin{matrix} -1\ 5 \\ \left\{ \begin{matrix} 0\ 2 \\ 1\ 1 \\ 2\ 2 \end{matrix} \right\} \\ 3\ 5 \end{matrix} \right\} \\ 4\ 10 \end{matrix} \right\} \text{ same outputs... —— That's the axis!}$$

The inputs/outputs are paired up as the latter start to repeat themselves – in reverse order – after we pass $x = 1$. There is, therefore, a *mirror symmetry* even though not at the y-axis...

Furthermore, we know that the "leading coefficient" a determined the large-scale behavior of the function:

- If $a > 0$, parabola opens up and there is a minimum.
- If $a < 0$, parabola opens down and there is a maximum.



Exercise 4.4.1

Describe such a function as a transformation.

Another familiar observation is that the slopes vary from location to location.

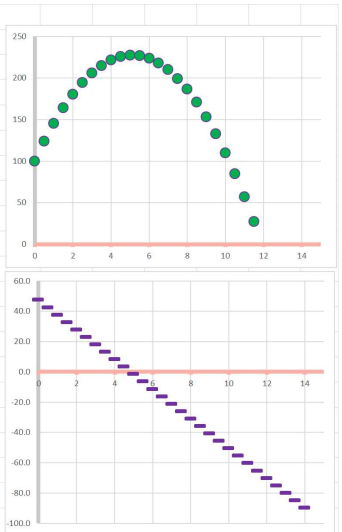
Example 4.4.2: shooting a cannon

Suppose a projectile is launched from a 100-meter tall building vertically up in the air with a speed of 50 meters per second with a gravity-caused acceleration of -9.8 meters per second squared. The altitude of the projectile is then modeled by a quadratic function:

$$y = f(x) = \frac{-9.8}{2}x^2 + 50x + 100\,.$$

We plot both the altitude (first plot) and the velocity of this projectile (second plot):

h=	0.50	time	altitude	velocity
	n	x	y	DQ
	1	0.0	100.0	
	2	0.5	123.8	47.6
	3	1.0	145.1	42.7
	4	1.5	164.0	37.8
	5	2.0	180.4	32.9
	6	2.5	194.4	28.0
	7	3.0	205.9	23.1
	8	3.5	215.0	18.2
	9	4.0	221.6	13.3
	10	4.5	225.8	8.3
	11	5.0	227.5	3.5
	12	5.5	226.8	-1.5
	13	6.0	223.6	-6.3
	14	6.5	218.0	-11.3
	15	7.0	209.9	-16.2
	16	7.5	199.4	-21.1
	17	8.0	186.4	-26.0
	18	8.5	171.0	-30.9
	19	9.0	153.1	-35.8



Of course, the velocity is found as the difference quotients of the altitude.

Exercise 4.4.3

(a) When does the projectile reach the highest point? (b) How fast does it go at that moment? (c) How fast does it hit the ground?

Exercise 4.4.4

(a) Plot the altitude and the velocity of a projectile launched out of a 10-meter deep trench with a speed of 20 meters per second. (b) Answer the questions asked in the last exercise.

The information we have collected previously allows us to easily answer the questions on our list:

1. Bounded or unbounded? Unbounded!
2. One-to-one or onto? No and No!
3. Increasing or decreasing? Increasing *then* decreasing or vice versa!
4. Odd or even? Even but only if we *shift* it!
5. Periodic? No!

Exercise 4.4.5

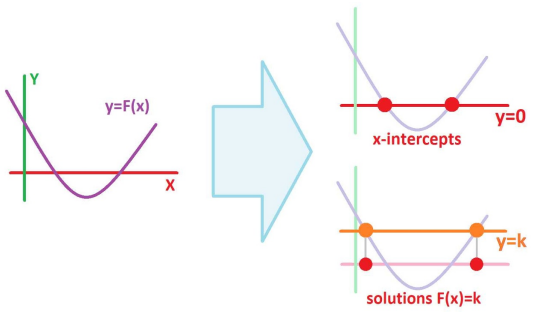
Justify and elaborate on the answers.

Below we will look at more subtle features of this function.

Recall that the x-intercepts of a function $y = f(x)$ are the solutions of the equation $f(x) = 0$. In the present context, we often use a different language.

Definition 4.4.6: roots of quadratic polynomial

The *roots* of a quadratic polynomial f are the solutions of the quadratic equation: $f(x) = 0$.



A familiar, and very important, formula is below.

Theorem 4.4.7: Quadratic Formula

The roots of a quadratic polynomial

$$f(x) = ax^2 + bx + c,\ a \neq 0\,,$$

are given by the following:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Exercise 4.4.8

State the theorem as an implication ("if ... then ..."). Is the converse true?

Exercise 4.4.9

Prove the theorem.

The following simple example shows the limitations of the applicability of the formula:

- The function $f(x) = x^2 + 1$ has no x-intercepts.

The formula is frequently written as follows:

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\,,$$

which means that there are (up to) two solutions:

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

We can imagine that the two come from the two inverses of the function...

Theorem 4.4.10: Vieta's Formulas

The roots x_1, x_2 of the quadratic polynomial $f(x) = ax^2 + bx + c$ satisfy the following equations:

$$x_1 + x_2 = -\frac{b}{a} \quad \text{and} \quad x_1 \cdot x_2 = \frac{c}{a}$$

Exercise 4.4.11

Prove the formulas.

A very important result below gives us a new form of the quadratic function.

Theorem 4.4.12: Factored Form of Quadratic Polynomial

If x_1, x_2 are the roots of a quadratic polynomial $f(x) = ax^2 + bx + c$, then we have:

$$f(x) = a(x - x_1)(x - x_2)$$

Proof.

We simply substitute and then notice that multiplication by 0 always produces 0:

$$f(x_1) = a(x - x_1)(x - x_2)\Big|_{x=x_1} = a(x_1 - x_1)(x_1 - x_2) = a \cdot 0 \cdot (x_1 - x_2) = 0\,.$$

$$f(x_2) = a(x - x_1)(x - x_2)\Big|_{x=x_2} = a(x_2 - x_1)(x_2 - x_2) = a \cdot (x_2 - x_1) \cdot 0 = 0\,.$$

Exercise 4.4.13

State the converse of the theorem. Is it true?

So, there may be two *linear* factors!

As the solutions to the equation

$$f(x) = ax^2 + bx + c = 0\,,$$

these numbers may or may not exist or they might coincide. The formula only produces x-intercepts when *what's inside the square root* is non-negative.

Definition 4.4.14: discriminant

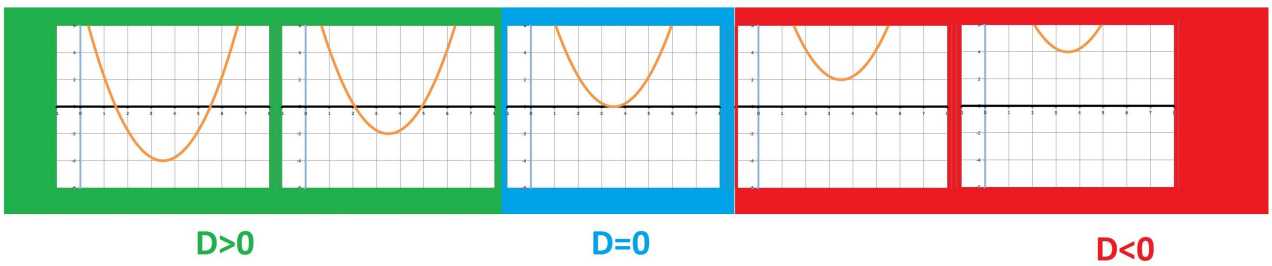
The *discriminant* of a quadratic polynomial

$$f(x) = ax^2 + bx + c,\ a \neq 0\,,$$

is the number defined by the following:

$$D = b^2 - 4ac$$

Now, let's see what decreasing the value of the discriminant D (which may be achieved by increasing the value of c, for example) does to the graph of $y = f(x)$. This process transforms the graph; it moves upward. Initially, the graph has two x-intercepts but when D, for a moment, reaches 0, the graph touches the x-axis. As D becomes negative, the graph passes the x-axis entirely so that the x-intercepts disappear:



In other words, the two x-intercepts start to get closer to each other, then merge, and finally disappear.

Warning!

The case $D = 0$ is a "borderline" case.

It makes sense to find the discriminant first and classify the equation:

Step 1:	Step 2:	Step 3:				
D	#		x_1	and	x_2	
$D > 0$	2	$-\frac{b}{2a} - \frac{\sqrt{D}}{2a}$		$\neq$	$-\frac{b}{2a} + \frac{\sqrt{D}}{2a}$	
$D = 0$	1		$-\frac{b}{2a}$	$=$	$-\frac{b}{2a}$	
$D < 0$	0			$\varnothing$		

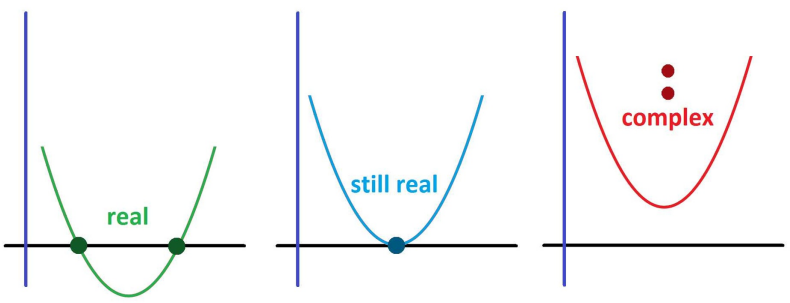
As a summary, we have the following important classification of quadratic polynomials in terms of *the sign of the discriminant*.

Theorem 4.4.15: x-Intercepts of Parabola

The two x-intercepts of a quadratic polynomial are:

- *distinct when its discriminant D is positive;*
- *equal when its discriminant D is zero;*
- *absent when its discriminant D is negative.*

We will see in Chapter 5DE-2 that the last case produces two *complex* roots:



The picture shows what happens as we are increasing the value of c.

Exercise 4.4.16

What will happen to the graph and to the x-intercepts as we increase the value of a? What about the value of b?

When the discriminant is zero, we say that the quadratic polynomial is a *complete square*; it just factors in this special way:

$$f(x) = a(x - x_1)^2 .$$

We also say that the factor, and the root, has *multiplicity* 2.

Definition 4.4.17: irreducible factor

When a quadratic polynomial cannot be factored into linear factors, it is called *irreducible*.

We conclude the following:

Theorem 4.4.18: Irreducible Quadratic Polynomial

When the discriminant is negative, the polynomial is irreducible.

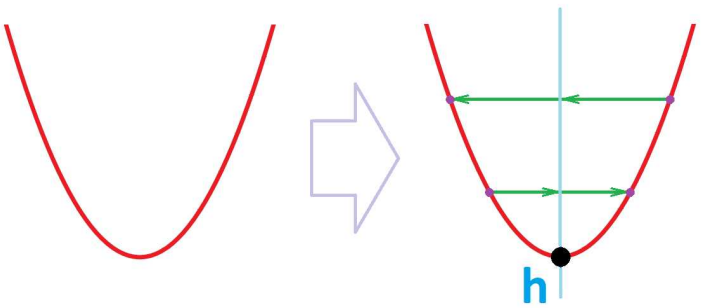
Now, we know that the axis of a parabola lies half-way between the two x-intercepts. Therefore, it is

$$h = \frac{x_1 + x_2}{2} = \frac{1}{2}\left(\frac{-b - \sqrt{b^2 - 4ac}}{2a} + \frac{-b + \sqrt{b^2 - 4ac}}{2a}\right) = -\frac{b}{2a}.$$

The equation

$$y = -\frac{b}{2a}$$

is the equation of the *axis* of the parabola:



Theorem 4.4.19: Maximum-Minimum of Quadratic Polynomials

Suppose $f(x) = ax^2 + bx + c$, $a \neq 0$, *is a quadratic polynomial. Then, the value*

$$h = -\frac{b}{2a}$$

is its extreme point. Moreover,

- *h is the global minimum point of f when $a > 0$, and*
- *h is the global maximum point of f when $a < 0$.*

Warning!

The formula (and the proof) works even when $D < 0$ and there are no x-intercepts!

This information can also come from completing a square (Chapter 3) and finding the *vertex form* of the quadratic function:

$$f(x) = a(x - h)^2 + k$$

Definition 4.4.20: vertex of parabola

The *vertex* of a parabola is the point on the graph that corresponds to the maximum or the minimum of the quadratic function.

It follows that the x-coordinate of the vertex of parabola is $h = -\frac{b}{2a}$.

Theorem 4.4.21: Range/Image of Quadratic Polynomial

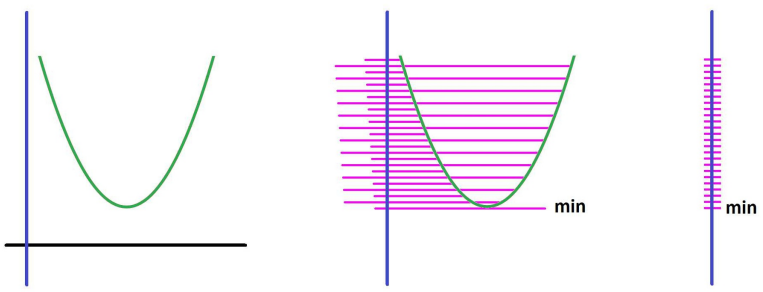
The range/image of a quadratic polynomial $y = ax^2 + bx + c$ is a closed ray:

- $[m, +\infty)$ *when* $a > 0$, *or*
- $(-\infty, M]$ *when* $a < 0$,

where m and M are the minimum and the maximum values of the function, respectively; they are equal to the value of f at the vertex:

$$m = f(h) \quad \text{or } M = f(h)\,.$$

The number is a bound of the image of the function:



Exercise 4.4.22

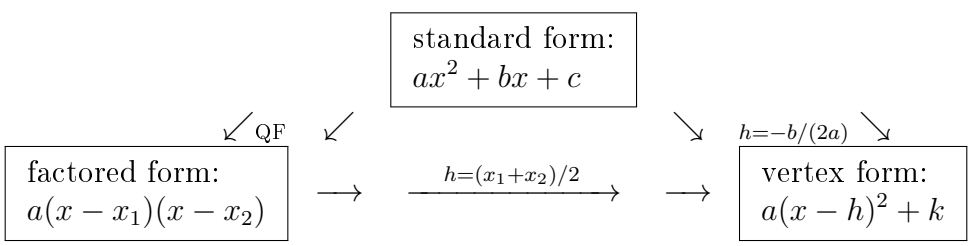
Is it possible for a quadratic polynomial to have exactly one linear factor?

Exercise 4.4.23

Simplify:

$$a\left(x - \frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(x - \frac{-b - \sqrt{b^2 - 4ac}}{2a}\right) = \,?$$

This is the summary of the three algebraic representations (forms) of quadratic polynomials and the transitions among them:



4.5. The polynomial functions

What do linear polynomials $f(x) = mx + b$ and quadratic polynomials $g(x) = ax^2 + bx + c$ have in common? There is no division!

The following concept will be routinely used.

Definition 4.5.1: polynomial

A *polynomial* is a (numerical) function computed from its input via addition, subtraction, and multiplication only.

No division – no chance of dividing by zero! We conclude the following:

Theorem 4.5.2: Domain of Polynomial

The domain of a polynomial is the set of all real numbers, $\mathbf{R} = (-\infty, +\infty)$.

Multiplication includes multiplication by a constant real number as well as by the input, x, itself. For example, we might have:

$$\begin{array}{lllllll} x & \to & \boxed{\text{add 2}} & \to & \boxed{\text{multiply by 2}} & \to & \boxed{\text{add 3}} \\ & & & & & & \downarrow \\ x & \to & & & & \to & \boxed{\text{multiply by } x} \to y \end{array}$$

This is the formula of this function:

$$f(x) = \Big[\big((x+2)\cdot 2\big) + 3\Big] \cdot x\,.$$

As an extreme case, it is possible that x might not even appear! Then, in addition to linear and quadratic polynomials, we discover a new class: *constant polynomials*. It is also clear that the polynomials include all *power functions*.

The way we learned to understand the linear, $f(x) = mx + b$, and the quadratic polynomials, $g(x) = ax^2 + bx + c$, was to derive many of their properties from the values of their parameters, or *coefficients*: m, b and a, b, c, respectively. There is a *standard* way to represent a polynomial. For a linear polynomial, it is its "slope-intercept form". For a quadratic polynomial, the formula $g(x) = ax^2 + bx + c$ is the standard form, as opposed to, say, the factored form: $g(x) = a(x - x_1)(x - x_2)$.

We try to approach all polynomials this way and put them in a similar, manageable form. For example, this is how we simplify the one above:

$$f(x) = \Big[\big((x+2)\cdot 2\big) + 3\Big] \cdot x = \big[(2x+4) + 3\big] \cdot x = \big[2x + 7\big] \cdot x = 2x^2 + 7x\,.$$

We can do the same with any polynomial, no matter how many steps are involved.

What is the *standard* way to represent a polynomial? Let's follow the treatment of the linear and quadratic polynomials and ask: What do they have in common? To make the pattern clear, let's add a constant to the list:

polynomials	terms
constant:	5
linear:	$2x + 5$
quadratic:	$3x^2 + 2x + 5$

Similar terms are aligned vertically! Accordingly, they have these special names, no matter in what polynomials they appear:

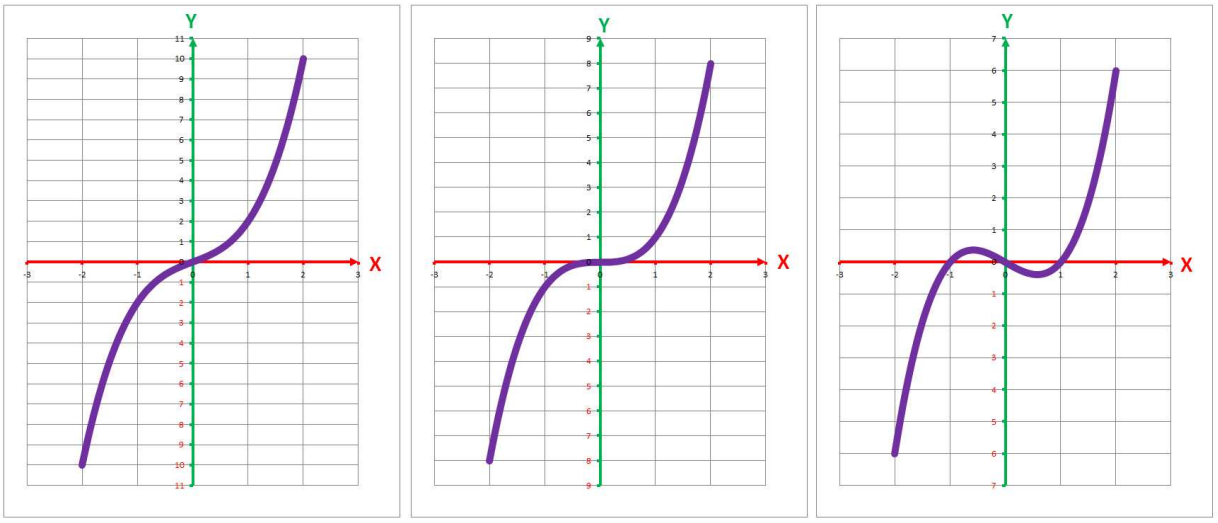
terms:	quadratic		linear		constant
	$3x^2$	$+$	$2x$	$+$	5

We add the cubic to the list. We can see the pattern even clearer if we identify the powers of x (keeping in mind that $x^0 = 1$):

constant:							$5x^0$
linear:					$2x^1$	$+$	$5x^0$
quadratic:			$3x^2$	$+$	$2x^1$	$+$	$5x^0$
cubic:	$-2x^3$	$+$	$3x^2$	$+$	$2x^1$	$+$	$5x^0$

The powers of x go down all the way to 0. Furthermore, we progress to the next line by just adding an extra term – a term of a higher power!

What does it do to the function? Just take a look at the graphs. All linear polynomials are lines and all quadratic polynomials are parabolas. Consider, for contrast, the extra degree of complexity – several distinctly different shapes of the graph – that the cubic polynomials might have:



Indeed, the number of ways the monotonicity may change has increased:

- linear: ↘ or ↗
- quadratic: ↘↗ or ↗↘
- cubic: ↘ or ↗ or ↘↗↘ or ↗↘↗

Exercise 4.5.3

Describe these functions as transformations.

Definition 4.5.4: degree of polynomial

The *degree* of a polynomial is the number n that is the highest power of x that needs to be computed. The degree of a polynomial P is denoted by

$$\deg P$$

From now on, we will prefer to identify polynomials according their degrees:

degrees	3		2		1		0	
0							$\boxed{5x^0}$	1 term
1					$\boxed{2x^1}$	$+$	$5x^0$	2 terms
2			$\boxed{3x^2}$	$+$	$2x^1$	$+$	$5x^0$	3 terms
3	$\boxed{-2x^3}$	$+$	$3x^2$	$+$	$2x^1$	$+$	$5x^0$	4 terms
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$

As you can see, the terms are also assigned degrees, instead of names.

Warning!

All terms can be zero except the leftmost.

Exercise 4.5.5

Use the Binomial Formula to find the standard form of $(2x+3)^6$. Show the correspondence between the triangle in the table above and the Pascal triangle.

So, the nth degree polynomial will be located in the $(n+1)$st row of the table. It will have $n+1$ terms, say:

$$f(x) = 2x^n - 11x^{n-1} + \ldots - 2x^3 + 3x^2 + 2x^1 + 5x^0\,.$$

Here "..." indicates the continuation of the pattern: declining, by 1, degrees of x. If we extract the coefficients from the formula, we realize that we have a finite *sequence* (Chapter 1):

$$2,\ -11,\ ...,\ -2,\ 3,\ 2,\ 5\,.$$

For a general formula, we choose letters to represent the coefficients of the powers of x with the subscripts indicating the power of the corresponding term:

$k=$	0	1	2	...	$n-1$	n
$a_k=$	a_0	a_1	a_2	...	a_{n-1}	a_n

For the above polynomial, we have:

the degrees of the terms:	n	$n-1$	...	3	2	1	0
the polynomial:	$2x^n$	$-11x^{n-1}$	...	$-2x^3$	$+3x^2$	$+2x^1$	$+5x^0$
the terms:	$2x^n$,	$-11x^{n-1}$,	...	$-2x^3$,	$3x^2$,	$2x^1$,	$5x^0$
the coefficients:	$a_n=2$,	$a_{n-1}=-11$,	...	$a_3=-2$,	$a_2=3$,	$a_1=2$,	$a_0=5$

Definition 4.5.6: standard form of polynomial

A polynomial in a *standard form* is given by a formula:

$$f(x)=a_nx^n+a_{n-1}x^{n-1}+...+a_2x^2+a_1x+a_0$$

where

$$a_n,\ a_{n-1},\ ...,\ a_2,\ a_1,\ a_0$$

are real numbers called the *coefficients* of the polynomial.

We can re-assemble the polynomial from its sequence of coefficients.

Exercise 4.5.7

Write the standard form of polynomial in the sigma notation.

Definition 4.5.8: terms of polynomial

The terms of a polynomial are named as follows:

$f(x)$				
	$=$	a_nx^n	nth	*leading term*
	$+$	$a_{n-1}x^{n-1}$	$(n-1)$th	
	...	...	...	
	$+$	a_2x^2	2nd	*quadratic term*
	$+$	a_1x^1	1st	*linear term*
	$+$	a_0	0th	*constant term*

The coefficients are named accordingly.

Of course, some of these coefficients can be zero. The exception is the leading one, a_n.

Theorem 4.5.9: Leading Term vs. Degree of Polynomial

The degree of a polynomial is the number n that is the highest power of x present in its standard form.

In other words, any polynomial with coefficients

$$a_0,\ a_1,\ a_2, ...,$$

satisfies:

$$\deg f = n \implies a_n \neq 0,\ a_{n+1} = 0,\ a_{n+2} = 0, \ldots$$

Exercise 4.5.10

What about the converse?

As we will see in Chapter 2DC-1, the leading term determines the *large-scale* behavior of the polynomial. Below, we are mostly concerned with the *small-scale* issues.

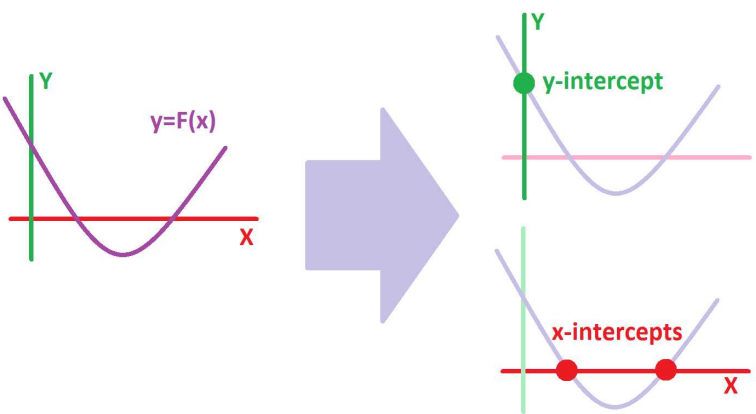
For example, what is the meaning of the constant term?

Theorem 4.5.11: Constant Term Is y-Intercept

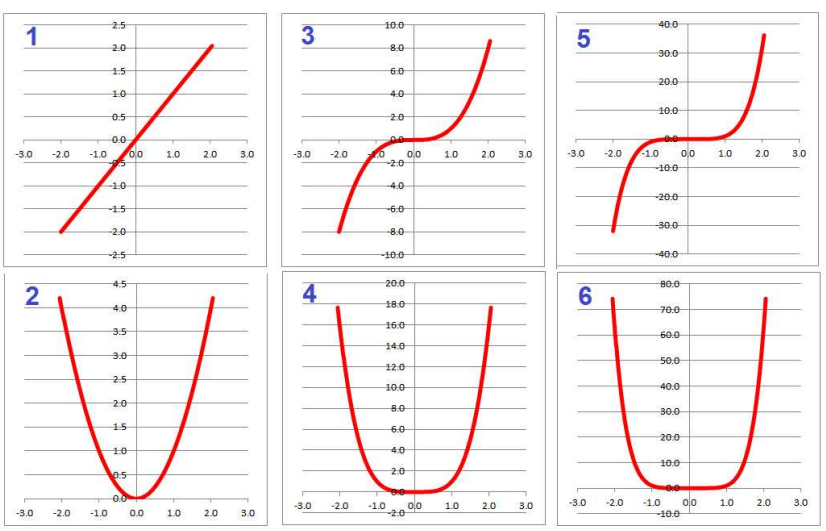
The constant term a_0 of a polynomial $f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_2 x^2 + a_1 x + a_0$ is its y-intercept.

Proof.

$$\begin{aligned} f(0) &= a_n x^n + a_{n-1}x^{n-1} + \ldots + a_2 x^2 + a_1 x + a_0 \Big|_{x=0} \\ &= a_n 0^n + a_{n-1}0^{n-1} + \ldots + a_2 0^2 + a_1 0 + a_0 \\ &= a_0 \,. \end{aligned}$$



So, in addition to the definition of polynomials in the beginning of this section, we can say that polynomials are "made" of powers of x (presented in Chapter 2):

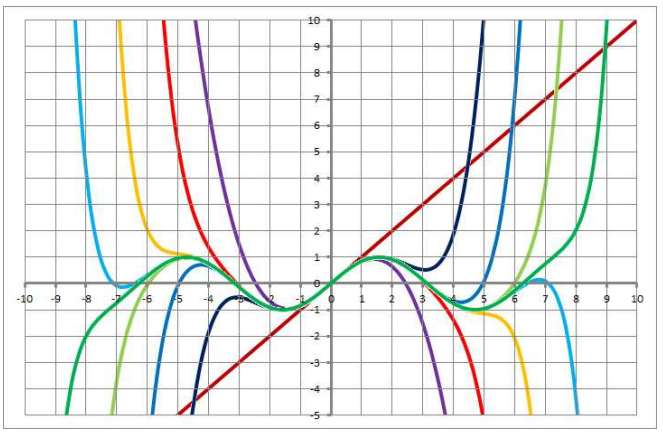


They are combined via multiplication by constant coefficients followed by addition (called "linear combina-

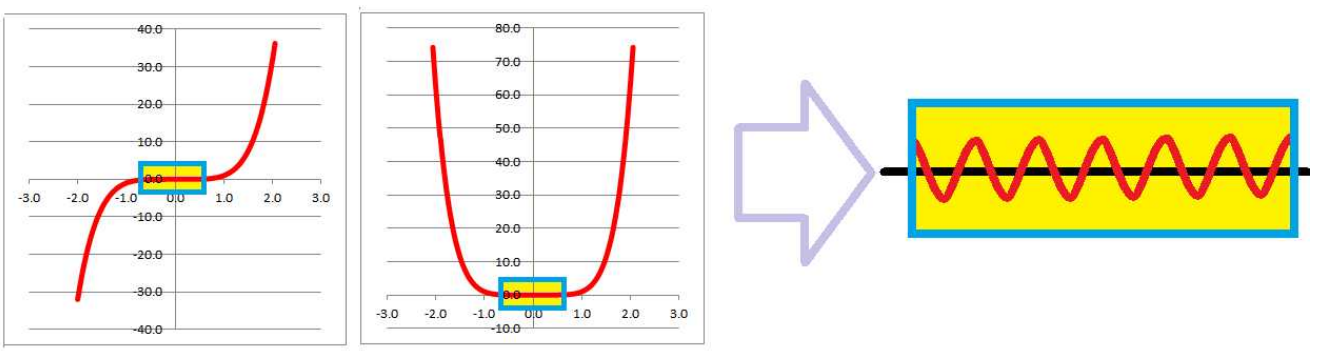
tions"):

$$\begin{array}{rcccccccc} & & x^3 & & x^2 & & x^2 & x^1 & x^0 = 1 \\ & & 2 \cdot x^3 & & 4 \cdot x^2 & & (-1) \cdot x^2 & 0 \cdot x^1 & 10 \cdot 1 \\ f(x) & = & 2x^3 & + & 4x^2 & - & x^2 & + & 10 \end{array}$$

It is the presence of various degree that explain the variety of behaviors of polynomials:



However, we will see that, when we zoom out from them, the graphs of polynomials resemble those of the power functions! On the other hand, when zoomed in on the x-axis, we might see many x-intercepts:



How many?

We have learned from our experience with linear and, especially, quadratic polynomials that there is another, just as important form for a polynomial: the *factored form*.

For example, a *linear polynomial* is factored as follows:

$$f(x) = mx + b = m\left(x + \frac{b}{m}\right), \ m \neq 0\,.$$

The factor in parentheses has a special importance because we can *read* something important directly from it:

- ▶ $x = -\dfrac{b}{m}$ is the x-intercept of f.

Next, suppose a *quadratic polynomial* happens to have two real roots x_1 and x_2 (they come from the Quadratic Formula). They give us two, or one if equal, x-intercepts and two (possibly equal) factors:

$$f(x) = ax^2 + bx + c = a(x - x_1)(x - x_2)\,.$$

We see how the standard form is converted to the factored form with two linear factors. Once again, the factors have a special importance because we can *read* something important directly from them:

- ▶ x_1 and x_2 are is the x-intercepts of f.

When the two x-intercepts are equal, $x_1 = x_2$, then so are the factors, and we have a complete square:

$$f(x) = a(x - x_1)(x - x_1) = a(x - x_1)^2\,.$$

Even though there are still two factors though, there is only one x-intercept!

A truly different situation is that of an irreducible quadratic polynomial, such as x^2+1; it has no x-intercepts nor linear factors! We say that there is one – irreducible quadratic – factor. What makes it "irreducible" is that fast that it cannot be reduced to factors of lower degree, i.e., linear factors.

Exercise 4.5.12

What is the degree of the product of two polynomials?

What do we know about *factorization of polynomials* in general?

One can take advantage of one's prior experience with integers (Chapter 1) by following these pairs of matching observations.

Analogy: integers vs. polynomials

Integers	Polynomials
The *sum* and the *product* of two integers is an integer.	The *sum* and the *product* of two polynomials is a polynomial.
A *prime* number cannot be further factored into smaller integers > 1. $2, 3, 5, 7, 11, 13, \ldots$	An *irreducible* polynomial cannot be further factored into lower degree polynomials. $x-1, x+2, x^2+1, \ldots$
Fundamental Theorem of Arithmetic: Every integer can be represented as the *product* of primes.	**Fundamental Theorem of Algebra:** Every polynomial can be represented as the *product* of irreducible factors.
This representation is *unique*, up to the order of the factors. $q = 2^5 \cdot 3^2 \cdot 5^1$ The *multiplicity* is how many times it appears.	This representation is *unique*, up to the order of the factors. $Q = (x-1)^5 \cdot (x+2)^2 \cdot (x^2+1)$ The *multiplicity* is how many times it appears.
	Irreducible polynomials have degrees 1 or 2.

Exercise 4.5.13

What is the degree of the polynomial Q above?

Corollary 4.5.14: Degree of Factored Polynomial

The sums of the degrees of the factors of a polynomial of degree n is n.

Exercise 4.5.15

Prove the corollary.

It is the purpose of the factored form to display the x-intercepts. The fact that we can see them follows from this fundamental result about numbers.

Theorem 4.5.16: Zero Factor Property

If the product of two numbers is zero, then either one of them or both is zero too. Conversely, if either one of the two numbers is zero, then so is their product.

In other words, we have the following:

$$a = 0 \quad \texttt{OR} \quad b = 0 \iff ab = 0$$

The concept of x-intercept refers to the geometry of the graph; the following is an algebraic substitute.

Definition 4.5.17: roots of polynomial

The *roots* of a polynomial P are the solutions of the polynomial equation:

$$P(x) = 0\,.$$

Then, substituting its root $x = x_1$ into the linear polynomial $L(x) = x - x_1$ makes it equal to zero – as well as any polynomial that has it as a factor! The following is crucial.

Theorem 4.5.18: Linear Factor Theorem

A linear factor $(x - x_1)$ is present in the factorization of a polynomial P if and only if $x = x_1$ is an x-intercept of P.

In other words, we have:

$$(x - x_1) \text{ is a factor of } P \iff P(x_1) = 0\,.$$

Warning!

Prove the $\Longrightarrow$ part of the theorem.

An important fact follows.

Corollary 4.5.19: Number of x-Intercepts

A polynomial cannot have more x-intercepts than its degree.

The theorem establishes a correspondence:

$$\begin{array}{rcl} \text{Algebra} & \longleftrightarrow & \text{Geometry} \\ \text{factor } (x - x_1) & \longleftrightarrow & x\text{-intercept } x = x_1 \end{array}$$

It's the same thing... but there is more!

Definition 4.5.20: multiplicity of factors and roots

1. The *multiplicity of a linear factor* of a polynomial P is how many times it appears in the factorization of P.
2. The *multiplicity of a root* and, concomitantly, the *multiplicity of an x-intercept* of P is that of the corresponding linear factor, i.e., how many times it appears in the factorization of P.

We have a more precise correspondence now:

$$\begin{array}{rcl} \text{Algebra} & \longleftrightarrow & \text{Geometry} \\ \text{factor } (x - x_1)^m & \longleftrightarrow & x\text{-intercept } x = x_1 \text{ with multiplicity at least } m \end{array}$$

Exercise 4.5.21

In how many ways can the graph to cross the x-axis?

We will assume below that someone has already factored the polynomial for us...

Example 4.5.22: factored polynomial

Consider the polynomial

$$Q(x) = x(x-1)^5 \cdot (x+2)^2 \cdot (x-.5) \cdot (x^2+1) \cdot \ldots$$

Here, we assume that the rest of the factors are irreducible. The polynomial has x-intercepts (roots) corresponding to each of its linear factors, as follows:

$P =$	x	$(x-1)^5 \cdot$	$(x+2)^2 \cdot$	$(x-.5) \cdot$	$(x^2+1) \cdot \ldots$
linear factors:	$(x-0)$	$(x-1)$	$(x+2)$	$(x-.5)$	
multiplicities:	1	5	2	1	
x-intercepts:	0	1	-2	.5	
points on the graph:	$(0,0)$	$(1,0)$	$(-2,0)$	$(.5,0)$	

The quadratic factor, and the rest, doesn't contribute anything because it remains positive for every x. If we arrange the values of x, we can see what the x-axis looks like with the four points of the graph shown:

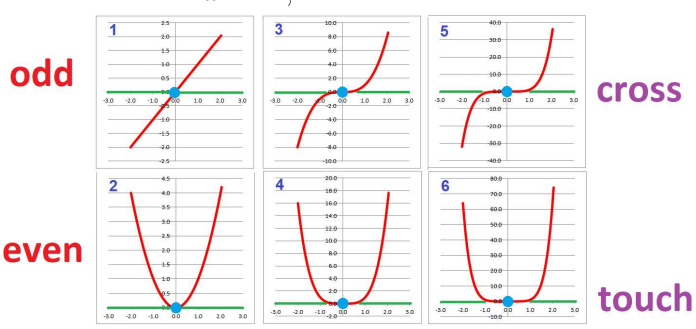
$$\begin{array}{rccccccc} x\text{-axis:} & - & \bullet & \bullet & \bullet & \bullet & \ldots & \longrightarrow x \\ & & -2 & 0 & .5 & 1 & \ldots & \end{array}$$

To be continued...

In other words, solving a polynomial equation $P(x) = 0$ and factoring the polynomial P is the same thing... except the former doesn't care about the *multiplicities*! But we do.

Example 4.5.23: power functions

Recall what we know about the power functions, x^1, x^2, x^3, ... (Chapter 2) but this time we also notice how differently the graphs intersect the x-axis; some cross and some touch:



But the exponents of these powers – $1, 2, 3, \ldots$ – are just the *multiplicities*! Of what? The factor $x = (x-0)$ and the root $x = 0$.

The example suggests the following convenient rule about the *nature of an x-intercept* as it is determined by its multiplicity.

Theorem 4.5.24: Multiplicity Rule for Polynomials

For every polynomial, we have the following:

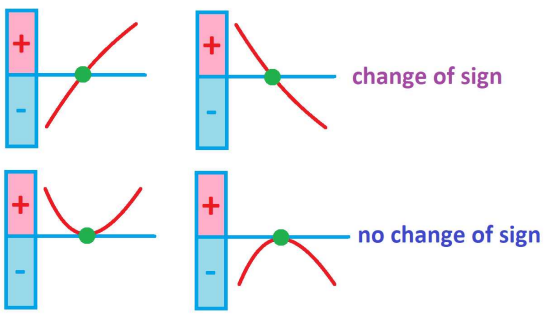
- *If the multiplicity of a linear factor is odd, the function takes both positive and negative values in the vicinity of the x-intercept; i.e., the graph crosses the x-axis: upward or downward.*
- *If the multiplicity of a linear factor is even, the function takes either only*

positive or only negative values in the vicinity of the x-intercept; i.e., the graph touches the x-axis: from above or from below.

Exercise 4.5.25

State the converse of either part of the theorem. Is it true?

So, we describe what happens in terms of the sign of y: it is positive in the upper half of the plane and negative in the lower half. Then, we use the language of "change of sign" of the function in the vicinity of the x-intercept:



Abbreviated, the theorem becomes the following convenient rule:

Linear factors and their x-intercepts

Algebra		**Geometry**			
multiplicity	sign	x-intercept			
odd	change of sign	↗	or	↘	cross
even	no change of sign	⌣	or	⌢	touch

Example 4.5.26: factored polynomial, continued

We continue with the last example. We assume that the rest of the factors of our polynomial are irreducible and positive:

$$Q(x) = x(x-1)^5 \cdot (x+2)^2 \cdot (x-.5) \cdot (x^2+1) \cdot \ldots$$

The x-intercepts and their multiplicities are listed below and they tell us how the graph interacts with the x-axis:

$Q(x) =$	$x\cdot$	$(x-1)^5\cdot$	$(x+2)^2\cdot$	$(x-.5)$	linear factors
x-intercepts:	0	1	-2	.5	read from the factorization
multiplicities:	1	5	2	1	read from the factorization
multiplicities are:	odd	odd	even	odd	read from the factorization
graph:	cross	cross	touch	cross	how it meets the x-axis
options:	↘	↘	⌣	↘	two choices per x-intercept
options:	↗	↗	⌢	↗	two choices per x-intercept

There are two possibilities for each x-intercept though. Does it mean that we have a total of 16 possibilities? No: if the graph crosses up, it can't cross up again. We have *two* possibilities:

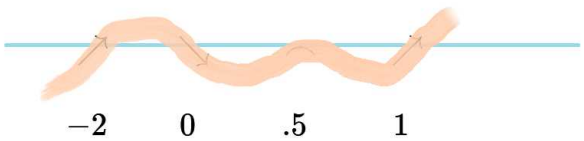
x-intercepts:	0	1	-2	.5
first option:	↘	↗	⌣	↘
second option:	↗	↘	⌢	↗

To find which one occurs, we pick a *starting point* for the graph; it serves as a "decider". How about $x = -10$? We substitute and discover that $Q(-10) < 0$. Therefore, there is a point to the left of the first x-intercept, $x = -2$, that lies *below* the x-axis. It follows that we must cross the axis *upward* at

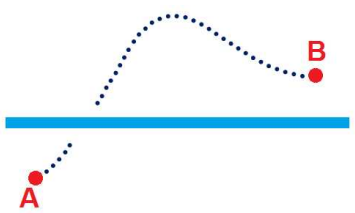
$x = -2$! Now we are *above* the x-axis. Therefore, we must cross the axis *downward* at $x = 0$! And so on. We order the x-intercepts, and then go from left to right, discovering which way we cross the axis at each x-intercept as soon as we reach it:

x-values:	-10		-2		0		$.5$		1	x-intercepts
y-values:	$-$		0	?	0	?	0	?	0	? above or below?
graph:			cross		cross		touch		cross	how it meets the x-axis
			$\nearrow$		$\searrow$		$\frown$		$\nearrow$	
y-values:	$-$		0	$+$	0	$-$	0	$-$	0	$+$ the answer

We can now draw a rough sketch of the graph:



Implicitly, we have relied on the following fact that one cannot cross a river from south to north twice in a row. To put it differently, if point A is on the southern bank of a river and point B on the northern bank, we'll have to swim:



We have the following fact (to be addressed in Chapter 2DC-2).

Theorem 4.5.27: Continuity of Polynomials

The graph of a polynomial can have a point below the x-axis and a point above it only if there is an x-intercept between them.

Exercise 4.5.28

State the theorem as an implication, an "if-then" statement.

Exercise 4.5.29

Apply the analysis in the last example to the polynomial

$$P(x) = x^2(2x-1)^3 \cdot (x-1)^3 \cdot (x^2-1)\,.$$

We can avoid using the *Multiplicity Rule* directly if we concentrate on the *signs* of the factors. We will base our analysis on a familiar table:

Rule of Signs for Multiplication

$$\begin{array}{rcl} (+) \cdot (+) & = & (+) \\ (+) \cdot (-) & = & (-) \\ (-) \cdot (+) & = & (-) \\ (-) \cdot (-) & = & (+) \end{array}$$

We will also use the theorem above that states that the sign of a polynomial can only change at an x-intercept.

Example 4.5.30: plotting polynomials

Let's analyze this polynomial

$$\begin{aligned} f(x) &= (x^4 - 9x^2)(x^2 - 4)^2(3x^2 + 2) && \text{We need to further factor it.} \\ &= x^2\big(x^2 - 3^2\big)(x^2 - 2^2)^2(3x^2 + 2) && \text{We use the Difference of Squares Formula.} \\ &= x^2(x-3)(x+3)(x-2)^2(x+2)^2 \ \cdot (3x^2 + 2)\,. \end{aligned}$$

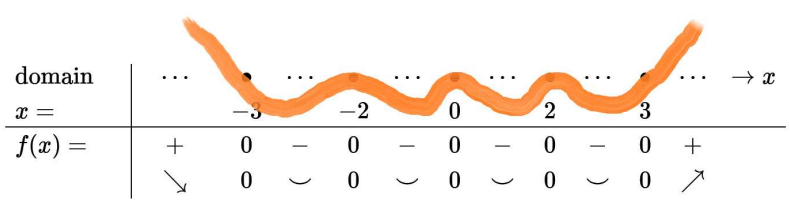
The x-intercepts are simply read off the list of linear factors:

$$x = 0,\ 3,\ -3,\ -2,\ 2\,.$$

This time, we will rely on the fact that at these points and at these points only can the function change its sign. We now list *all* the factors. They are simple enough for us to determine where and whether each changes its sign:

the points		-3		-2		0		2		3		
factors	signs											
x^2	$+$	$+$	$+$	$+$	$+$	0	$+$	$+$	$+$	$+$	$+$	
$x-3$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	0	$+$	
$x+3$	$-$	0	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	
$(x-2)^2$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	0	$+$	
$(x+2)^2$	$+$	0	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	
domain	$\cdots$	$\bullet$	$\cdots$	$\bullet$	$\cdots$	$\bullet$	$\cdots$	$\bullet$	$\cdots$	$\bullet$	$\cdots$	$\to x$
$x =$		-3		-2		0		2		3		
$f(x) =$	$+$	0	$-$	0	$-$	0	$-$	0	$-$	0	$+$	
	$\searrow$	0	$\smile$	0	$\smile$	0	$\smile$	0	$\smile$	0	$\nearrow$	

Here we go vertically in each column and determine the sign of the function using the above rule. An idea of what the graph of f looks like is seen in the last row: It crosses the x-axis downward, touches it three times from below, and then crosses it upward. We can sketch the graph based on the last row only:



The table solves for us this equation and these inequalities:

$$f(x) = 0,\ f(x) > 0,\ f(x) \geq 0,\ f(x) < 0,\ f(x) \leq 0\,.$$

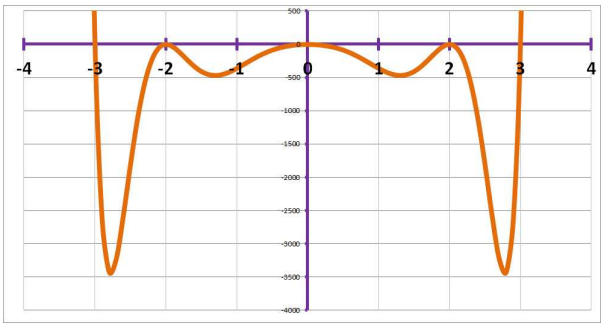
For example, the solution set of the inequality:

$$f(x) > 0, \quad \text{or} \quad (x^4 - 9x^2)(x^2 - 4)^2(3x^2 + 2) > 0\,,$$

is

$$(-\infty, -3) \cup (3, +\infty)\,.$$

We confirm our analysis with a graphic utility:



Exercise 4.5.31

Repeat the analysis but use the y-intercept as the "decider".

Exercise 4.5.32

Redo the problem using the Multiplicity Rule.

Exercise 4.5.33

Solve $f(x) \leq 0$.

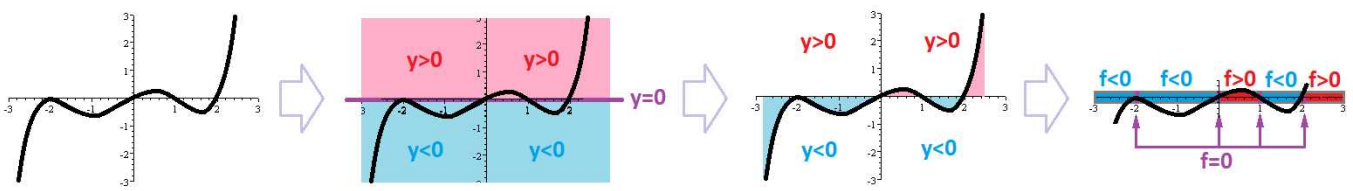
Exercise 4.5.34

Apply the analysis to the polynomial

$$f(x) = x^3(2x-2)\cdot(x-1)^2\cdot(x^2+2x+1)\,.$$

Example 4.5.35: inequalities as byproduct

This is what happens when one solves a polynomial equation or inequality when the polynomial is already factored; we have it plotted and the answers are read from the graph:



For example, the solution set of the equation:

$$f(x) = 0\,,$$

is

$$\{-2, 0, 1, 2\}\,.$$

The solution set of the inequality:

$$f(x) \leq 0\,,$$

is (in blue)

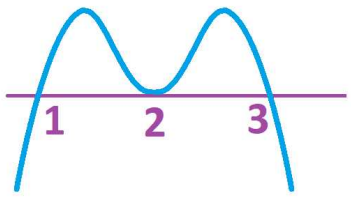
$$(-\infty, 0] \cup [1, 2]\,.$$

Exercise 4.5.36

Solve $f(x) > 0$.

Example 4.5.37: reconstruct function from graph

Let's find a plausible formula for the function the graph of which is sketched below:



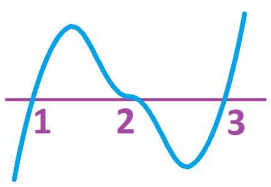
First, there are three x-intercepts: $1, 2, 3$. Their multiplicities are, respectively, odd, even, odd. The simplest choices of the multiplicities are $1, 2, 1$. Then the factors are $(x-1), (x-2), (x-3)$. There are no other linear factors! Then our function could be at its simplest the following:

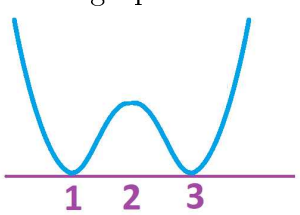
$$f(x) = -(x-1)(x-2)(x-3)\,.$$

The negative sign comes from the fact that the graph opens *down*.

Exercise 4.5.38

Suggest a plausible formula for each of the functions the graphs of which are sketched below:





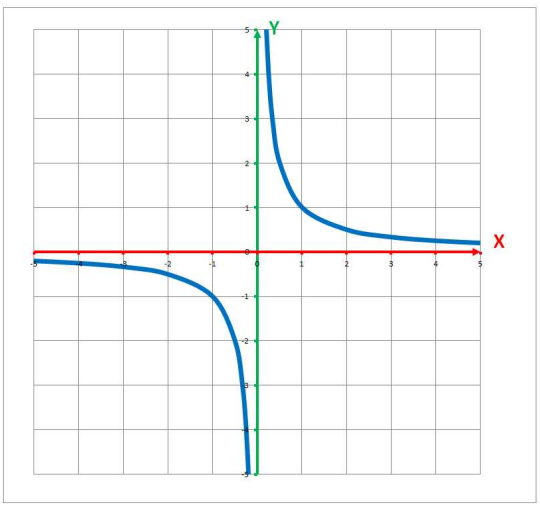
4.6. The rational functions

Example 4.6.1: reciprocal function

We are already familiar with such an important function as the reciprocal:

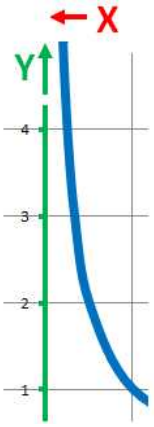
$$y = \frac{1}{x}\,.$$

Even with such a simple formula, once division is introduced, the complexity increases dramatically. We can see – in comparison to the polynomials – some new features in the graph:



First, *holes in the domain* appear. As we see above, the hole in the domain corresponds to a vertical

line on the xy-plane that the graph can't intersect! As the curve approaches this line, it has to start to *climb*, faster and faster, up or down:

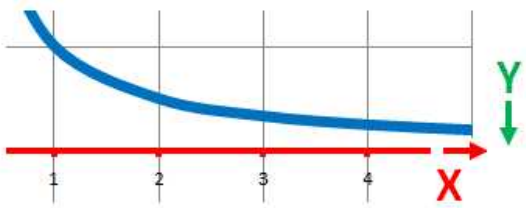


The result is a graph with a virtually vertical part close to this line. The phenomenon is also seen in the data:

x	1	1/2	1/3	...	1/100	...
$1/x$	1	2	3	...	100	...

Lines like this are called *vertical asymptotes*.

An opposite – but also very similar – behavior is seen along a certain horizontal line. If the graph can't cross it, it has to start to *crawl*, with virtually no up or down progress:



The phenomenon is also seen in the data:

x	1	2	3	...	100	...
$1/x$	1	1/2	1/3	...	1/100	...

Such a line is called a *horizontal asymptote*.

If we zoom out, the ends of the graph merge with the asymptotes. (The topic of asymptotes is fully addressed in Chapter 2DC-2.)

Exercise 4.6.2

Describe the symmetry alluded to in the example.

Exercise 4.6.3

Describe this function as a transformation.

We are adding *division* to the allowed operations now!

We use just integer numbers initially, but once we start dividing, we have to face *fractions*:

$$15 \text{ and } 23 \text{ produce } \frac{15}{23}.$$

Similarly, we use just polynomials initially, but once we start dividing, we have to face *fractions*:

$$x^2-1 \text{ and } x^3-x \text{ produce } \frac{x^2-1}{x^3-x}.$$

Another word for "fraction" is "ratio". That is why we use the following terminology:

- The fractions of integers are called *rational numbers*.
- The fractions of polynomials are called *rational functions*.

The following will be routinely used.

Definition 4.6.4: rational function

The ratio of two polynomials is called a *rational function*:

$$f(x) = \frac{P(x)}{Q(x)}\,.$$

Warning!

All polynomials are rational functions too.

The possibility of division makes the issue of the implied domain, in contrast to polynomial, a non-trivial matter. We factor the denominator and use the *Linear Factor Theorem* from the last section.

Example 4.6.5: domain of rational function

What is the domain of

$$f(x) = \frac{x-1}{x+1}\ ?$$

We look at the denominator, set it equal to zero, $x+1=0$, and solve for x. Then, $x=-1$ is not in the domain and the rest of the real numbers are.

Exercise 4.6.6

What is the domain of this function:

$$f(x) = \frac{x+1}{x+1}\ ?$$

Example 4.6.7: more domains of rational functions

What is the domain of

$$f(x) = \frac{x-1}{x^2-4}\ ?$$

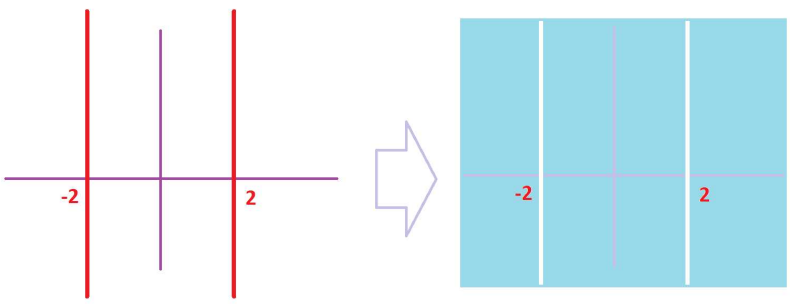
Set the denominator to 0 and solve:

$$\begin{aligned} x^2-4 &= 0 \quad \Longleftrightarrow \\ x^2 &= 4 \quad \Longleftrightarrow \\ x &= \pm 2\,. \end{aligned}$$

So, the domain consists of all real numbers except ± 2:

$$\text{domain } = \{x:\ x \neq \pm 2\} = (-\infty,-2)\cup(-2,2)\cup(2,+\infty)\,.$$

The solutions that we have found, $x=-2$ and $x=2$ are, in fact, the equations of two vertical lines that cut the xy-plane so that the graph can't cross them! Therefore, the graph will have to have three *branches* – each within one of these three parts of the plane:

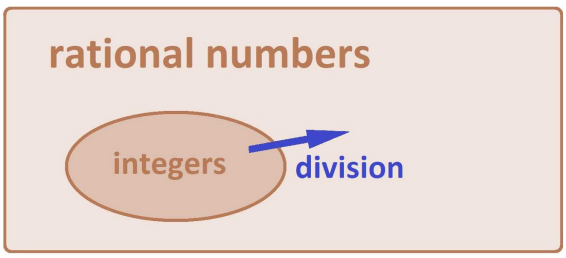


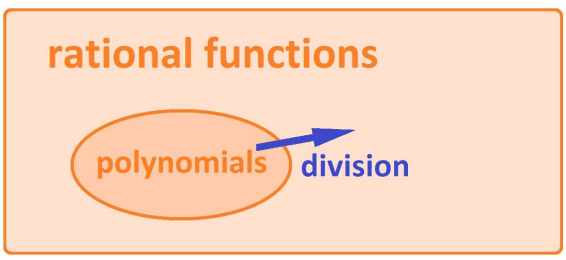
Exercise 4.6.8

What is the domain of

$$f(x) = \frac{x^2 - 1}{(x^2 + 4x + 4)(x^4 + 2)} \ ?$$

So, we discovered in the last section that the algebra of polynomials mimics the algebra of integers. We take another step in that direction and observe that the rational functions appear from polynomials – just like the rational numbers from integers – via division:





We now continue with the *analogy* we started in the last section.

Analogy: integers vs. polynomials

Integers	Polynomials
The *sum* and the *product* of two integers is an integer.	The *sum* and the *product* of two polynomials is a polynomial.
The *ratio* of two integers is called a rational number.	The *ratio* of two polynomials is called a rational function.
Rational numbers	**Rational functions**
The sum, the product, and the ratio of two rational numbers is a rational number. $\frac{2}{3} + \frac{1}{2} = \frac{7}{3 \cdot 2}$	The sum, the product, and the ratio of two rational functions is a rational function. $\frac{x}{x-1} + \frac{1}{x+1} = \frac{x^2 + 2x - 1}{(x-1)(x+1)}$
The numerator and the denominator of every rational number can be represented as the products of different primes. $\frac{p}{q} = \frac{2^5 \cdot 5^2 \cdot 7^1 \cdot \ldots}{3^2 \cdot 11^2 \cdot \ldots}$	The numerator and the denominator of every rational function can be represented as the products of different irreducible factors. $\frac{P}{Q} = \frac{(x-1)^5 \cdot (x+2)^2 \cdot (x^2+1) \cdot \ldots}{(x+1)^2 \cdot (x-3)^1 \cdot (2x^2+1) \cdot \ldots}$

Warning!

Canceling factors will simplify a rational function, but it might also change its domain.

How do we find the x-intercepts of a rational function? Just as in the last section, we use the *Linear Factor Theorem* that states that substituting $x = x_1$ into the linear polynomial $x - x_1$ makes it equal to zero – as well as any polynomial that has it as a factor.

Example 4.6.9: x-intercepts of rational functions

Once both polynomials are factored, we just need to watch for possible cancellations:

$$\frac{x(x-1)}{(x+1)(x-1)} = \frac{x}{x+1} \quad \text{for} \quad x \neq 1\,.$$

So, $x_1 = -1$ is an x-intercept but $x_2 = 1$ is not!

This analysis implies the following:

Theorem 4.6.10: Rational Function's x-Intercepts

A rational function's x-intercept

- *is a root of its numerator, provided it*
- *isn't also among the roots of the denominator.*

Proof.

$$P(x) = 0 \quad \texttt{AND} \quad Q(x) \neq 0 \implies \frac{P(x)}{Q(x)} = 0\,.$$

Exercise 4.6.11

State the theorem in the form of an equivalence.

Warning!

Below we assume that the numerator and the denominator don't share factors.

The challenge of understanding rational functions is that the holes in the domain break the graph into *separate branches*.

Example 4.6.12: plot rational function

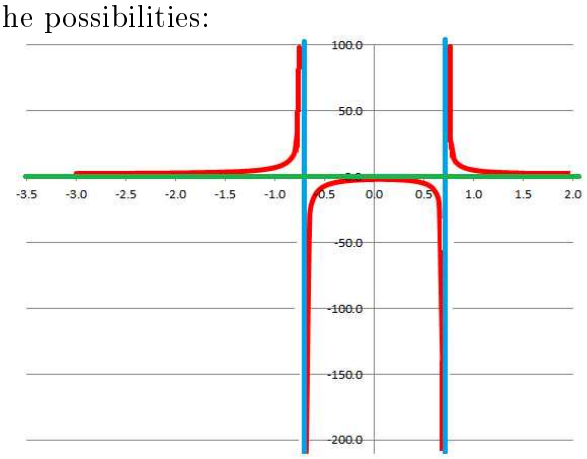
What does the plot of this function look like:

$$f(x) = \frac{x}{x^2 - 1}\ ?$$

We look at the denominator, set it equal to zero, $x^2 - 1 = 0$, and solve for x. Then, $x = \pm 1$ is not in the domain and the rest of the real numbers are. We then draw these vertical lines on the xy-plane:

$$x = -1 \quad \text{and} \quad x = 1\,.$$

The graph can't cross these lines! There are, therefore, three branches. Based on the available information, this is one of the possibilities:



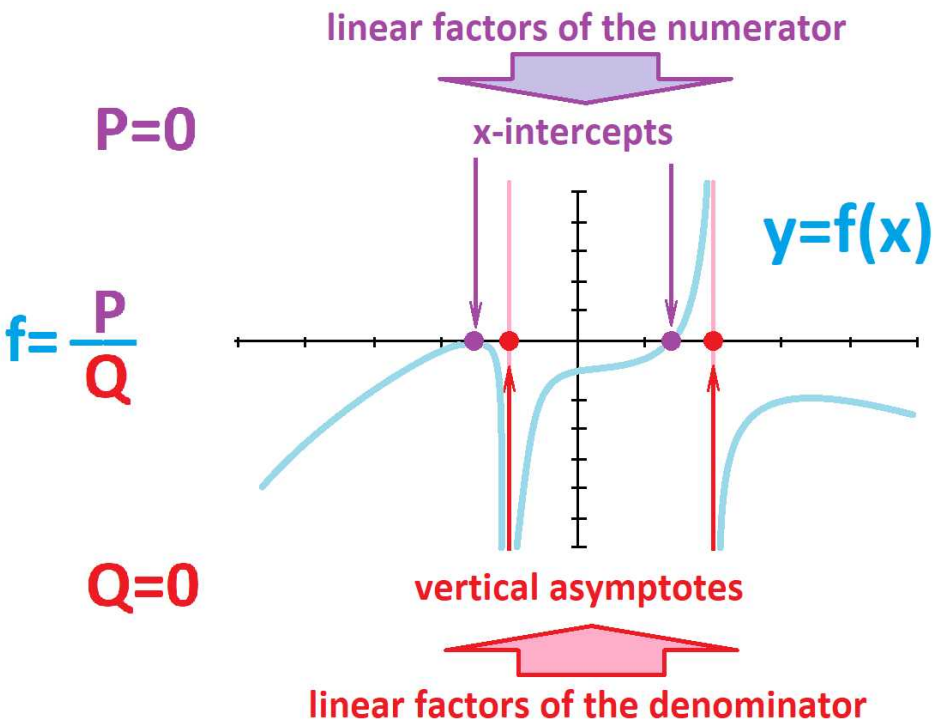
Exercise 4.6.13

Present other examples of graphs of functions with this domain and this x-intercept. Make up your own example. Repeat.

This is the summary of where the information about the graph of a rational function comes from:

Linear factor $(x - x_1)$	in numerator	$\rightarrow$	$x = x_1$	is an x-intercept.	$- - \bullet - -$
Linear factor $(x - x_2)$	in denominator	$\rightarrow$	$x = x_2$	is a vertical asymptote.	$- - \circ - -$

Multiplying by 0 and dividing by 0 produce very different results:



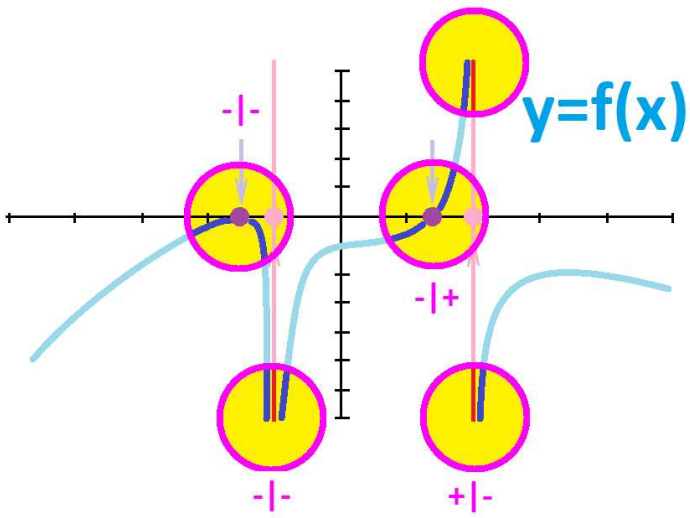
Warning!

A rational function doesn't have to have asymptotes, vertical or horizontal.

Just as in the case of polynomials, we have to look at the *multiplicities* of these linear factors. The effect of the linear factors of the *numerator* remains the same as in the last section; it produces x-intercepts with the following effect:

- The function is changing the sign (and the graph is crossing the x-axis) when the multiplicity is odd.
- The function is not changing the sign (and the graph is touching the x-axis) when the multiplicity is even.

The case of a linear factor in the *denominator* is very similar as we can continue to watch how the signs change. No change on the left, change on the right:



The difference is that the signs change not by crossing the x-axis but by (infinite) jumping.

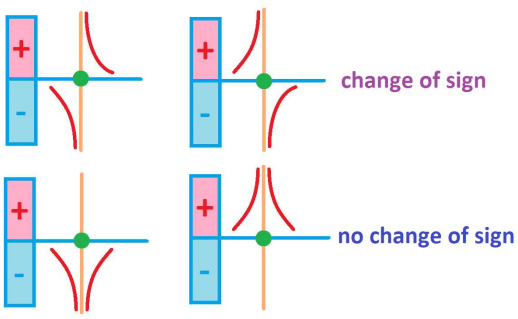
We summarize the result below.

Theorem 4.6.14: Multiplicity Rule for Rational Functions

For a linear factor in the denominator of a rational function, we have the following in the vicinity of the corresponding value of x:

- *If the multiplicity of the factor is odd, the graph of the function jumps over the x-axis: upward or downward.*
- *If the multiplicity of the factor is even, the graph of the function stays on the same side of the x-axis: above or below.*

We interpret the theorem in terms of the *change of sign*:



Abbreviated, the theorem becomes the following convenient rule:

Linear factors of denominator and asymptotes

Algebra		**Geometry**
multiplicity	sign	vertical asymptote
odd	change of sign	$\nearrow \vdots \nearrow$ or $\searrow \vdots \searrow$
even	no change of sign	$\nearrow \vdots \searrow$ or $\searrow \vdots \nearrow$

In contrast to the case of a polynomial, the jump over a vertical asymptote might be *infinite*!

Example 4.6.15: graph of rational function

Consider the rational function

$$f(x) = \frac{x^2 \cdot (x+2) \cdot (x-1) \cdot (x^2+1)}{(x+1) \cdot (x-3)^2 \cdot (x^2+2)}.$$

We analyze these two polynomials separately and derive two separate sets of information about the

function:

from the numerator...	x-intercepts:	$x =$	-2		0	1	
from the denominator...	vertical asymptotes:	$x =$		-1			3

The two irreducible quadratic factors don't contribute anything. This is what the x-axis looks like with the three points of the graph and two holes in the domain shown:

$$\begin{array}{rccccccccc} x\text{-axis:} & - & \bullet & \circ & \bullet & \bullet & \circ & \ldots & \longrightarrow & x \\ & & -2 & -1 & 0 & 1 & 3 & & & \end{array}$$

The numerator's roots and their multiplicities tell us whether the graph *passes across* the x-axis or stays on the same side:

x-intercepts:	-2	0	1
multiplicities:	1	2	1
graph:	cross	touch	cross

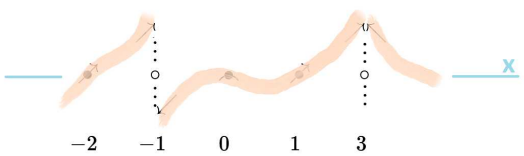
The denominator's roots and their multiplicities tell us whether the graph *jumps over* the x-axis or stays on the same side:

asymptotes:	-1	3
multiplicities:	1	2
graph:	jump	stay

There are still two possibilities for each of these points. Just as in the last section, we need a *starting point* as a "decider". We choose $x = -10$. Since $f(-10) < 0$, we have a point to the left of the left-most x-intercept, $x = -2$, that lies *below* the x-axis. It follows that we must cross the x-axis *upward* at $x = -2$. After that, we are *above* the x-axis and the process continues:

x-values:	-10	-2	-1	0	1	3	
graph:	$-$	$\bullet$	$\circ$	$\bullet$	$\bullet$	$\circ$	$\longrightarrow x$
graph:		cross	jump	touch	cross	stay	
$x-$axis:		$\nearrow$	$\vdots$ $\nearrow$	$\frown$	$\nearrow$	$\vdots$ $\nearrow\!\searrow$	$\longrightarrow x$

The bottom of the table is meant to visualize the data that we have collected. Furthermore, here is a rough sketch of the graph based on this data:



We can also derive solutions of the inequalities for the function from this sketch. For example, for the inequality

$$f(x) > 0\,,$$

the solution set is

$$(-2,-1) \cup (1,3) \cup (3,+\infty)\,.$$

And so on.

Unlike polynomials, rational functions *can* jump over the x-axis. However, it can only happen under very specific circumstances: vertical asymptotes. Implicitly, we have relied on the following fact (to be addressed in Chapter 2DC-2).

Theorem 4.6.16: Continuity of Rational Functions

The graph of a rational function can have a point below the x-axis and a point above the x-axis at the same time only if there is an x-intercept or a vertical asymptote between them.

Exercise 4.6.17

State the theorem as an implication, an "if-then" statement.

Exercise 4.6.18

Write the solution sets for the inequalities with f from the last example:

$$f(x) \geq 0,\ \ f(x) < 0,\ \ f(x) \geq 0\,.$$

We can avoid using the Multiplicity Rule directly if we concentrate on the *signs* of the factors. We will base our analysis on the following rules for division that are identical to those for multiplication used in the last section:

Rule of Signs for Division

$$\begin{array}{rcrcl} (+) & / & (+) & = & (+) \\ (+) & / & (-) & = & (-) \\ (-) & / & (+) & = & (-) \\ (-) & / & (-) & = & (+) \end{array}$$

Indeed, weather we multiply or divide, the sign of the term affects the result the same. We will also use the fact that the sign of a polynomial can only change at an x-intercept.

Example 4.6.19: another graph of rational function

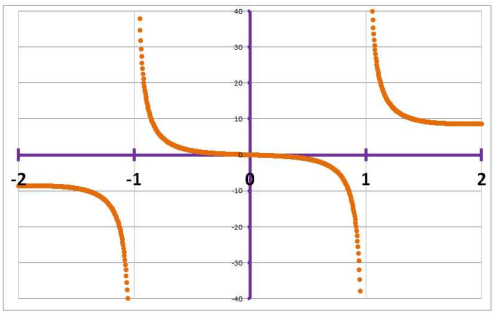
Let's analyze this rational function

$$f(x) = \frac{x(3x^2+1)}{(x-1)(x+1)}\,.$$

First, the domain is all reals except $x = \pm 1$. We now need to find the signs of the factors (dependent on x) and then the sign of the function. We take note of the domain and list all the factors present. We then determine where each of them is equal to zero and how its sign changes with x as it passes this location. Finally, we find the sign of $f(x)$ as it varies with x, one interval at a time using the *Rule of Signs* (for multiplication and for division) presented above. These are the results:

the points		-1		0		1		
factors	signs							
x	$-$	$-$	$-$	0	$+$	$+$	$+$	
$3x^2+1$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	
$x-1$	$-$	$-$	$-$	$-$	$-$	0	$+$	
$x+1$	$-$	0	$+$	$+$	$+$	$+$	$+$	
x-axis		$\circ$		$\bullet$		$\circ$		$\longrightarrow x$
$x=$		-1		0		1		
$f(x)$	$-$	$\vdots$	$+$	0	$-$	$\vdots$	$+$	

We confirm the result by using a graph-plotting application:



We can also derive solutions of the inequalities from the table. For example, for the inequality

$$f(x) \leq 0\,,$$

the solution set is

$$(-\infty, -1) \cup [0, 1)\,.$$

And so on.

Exercise 4.6.20

Solve $f(x) > 0$.

Exercise 4.6.21

Repeat the analysis but use the y-intercept as the "decider".

Exercise 4.6.22

Redo the problem using the Multiplicity Rule.

Exercise 4.6.23

Describe such a function as a transformation.

Exercise 4.6.24

Apply the analysis to the rational inequality:

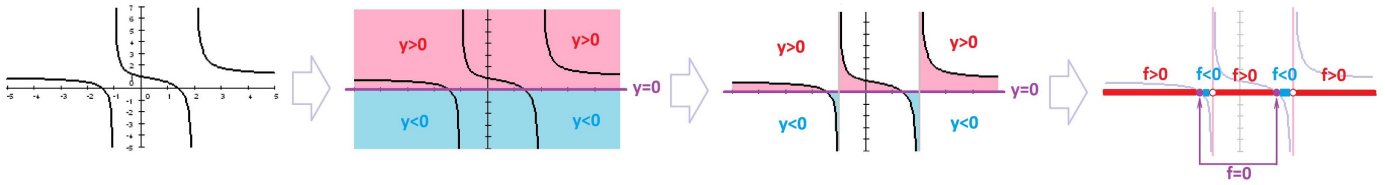
$$f(x) = \frac{x^3(2x-2)}{(x-1)^2 \cdot (x^2+2x+1)}\,.$$

As we look at the graphs, we realize that the effect of a linear factor in the numerator and the denominator of a rational function is, in a sense, *opposite*:

linear factor $(x - x_1)$	in numerator	$\rightarrow$	value of y at $x = x_1$	is 0	(x-intercept).
linear factor $(x - x_2)$	in denominator	$\rightarrow$	value of y at $x = x_2$	is ∞	(vertical asymptote).

Example 4.6.25: rational inequality as byproduct

This is what happens when we solve a *rational inequality* (and the corresponding equation) when the function comes already factored:



For example, the solution set of the equation:

$$f(x) = 0\,,$$

is

$$\{-1.5, 1.5\}\,.$$

The solution set of the inequality:

$$f(x) \leq 0$$

is

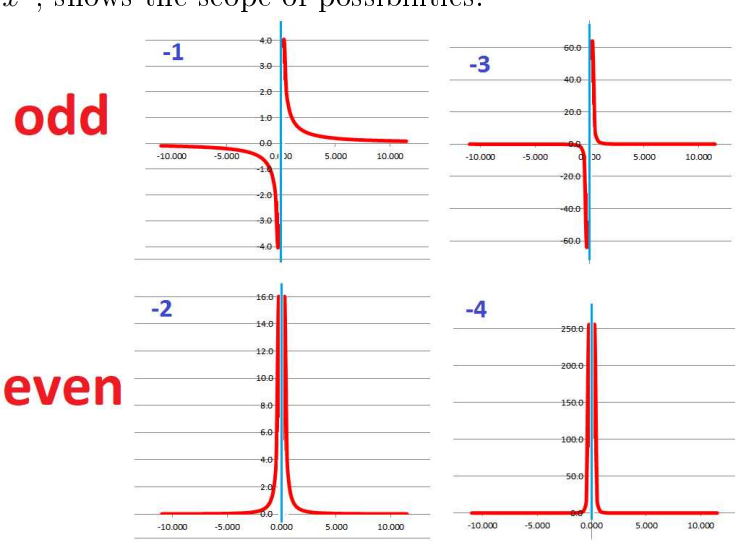
$$[-1.5, -1) \cup [1.5, 2)\,.$$

Exercise 4.6.26

Solve $f(x) > 0$.

Example 4.6.27: range of rational function

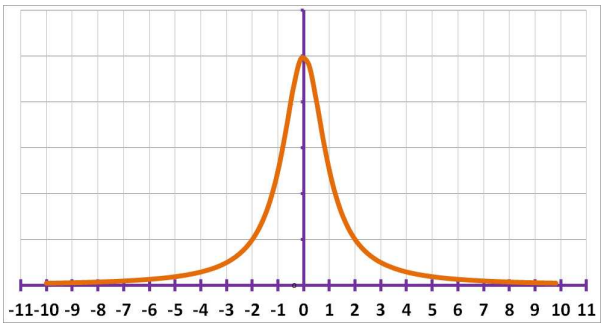
We have learned that its domain is entirely determined by the linear factors of the denominator of the rational function, $f(x) = P(x)/Q(x)$. What about the *range*? The example of the negative powers, i.e., $P(x) = 1,\ Q(x) = x^n$, shows the scope of possibilities:



So, we have:

- If n is odd, the range of $\dfrac{1}{x^n}$ is $(-\infty, 0) \cup (0, +\infty)$.
- If n is even, the range of $\dfrac{1}{x^n}$ is $(0, +\infty)$.

The example of $f(x) = \dfrac{1}{x^2+1}$ below shows that the range doesn't have to be infinite:

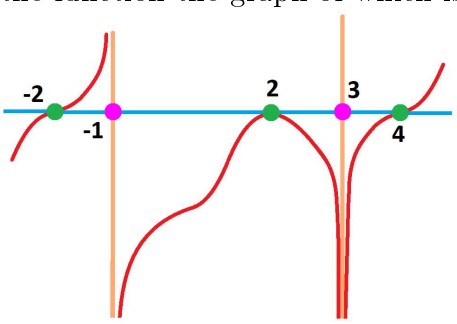


Exercise 4.6.28

Find the range of the last function and then describe the function as a transformation.

Example 4.6.29: reconstruct function from graph

Let's find a plausible formula for the function the graph of which is sketched below:



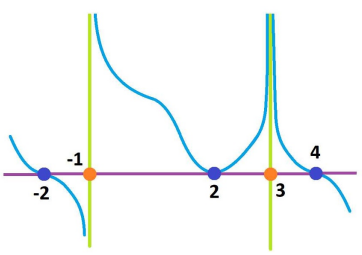
First, there are three x-intercepts: $x = -2, 2, 4$ with multiplicities, respectively, odd, even, odd. Then the factors for the numerator are chosen to be: $(x+2), (x-2)^2, (x-4)$. Second, there are two vertical asymptotes: $x = -1, 3$ with multiplicities, respectively, odd, even. Then the factors for the denominator are chosen to be: $(x+1), (x-3)^2$. Finally, our function could be at its simplest the following:

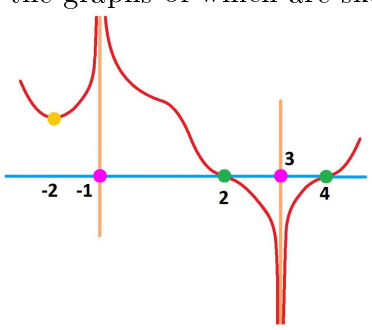
$$f(x) = \frac{(x+2)(x-2)^2(x-4)}{(x+1)(x-3)^2}.$$

The positive sign comes from the fact that the graph end at $+\infty$.

Exercise 4.6.30

Suggest a plausible formula for each of the functions the graphs of which are sketched below:





4.7. The root functions

Example 4.7.1: equations with powers

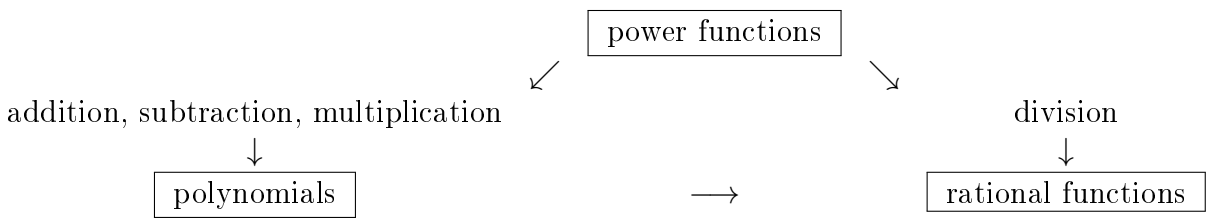
1. What should be the side of a square pool with an area of 200 square feet? We need to solve the equation:
$$x^2 = 200\,.$$
2. What should be the side of a cubic tank to contain 200 cubic feet of water? We need to solve the equation:
$$x^3 = 200\,.$$

If we are to do this repetitively, we'll need to figure out the inverses of these two power functions.

We have produced new functions starting from *power functions*,

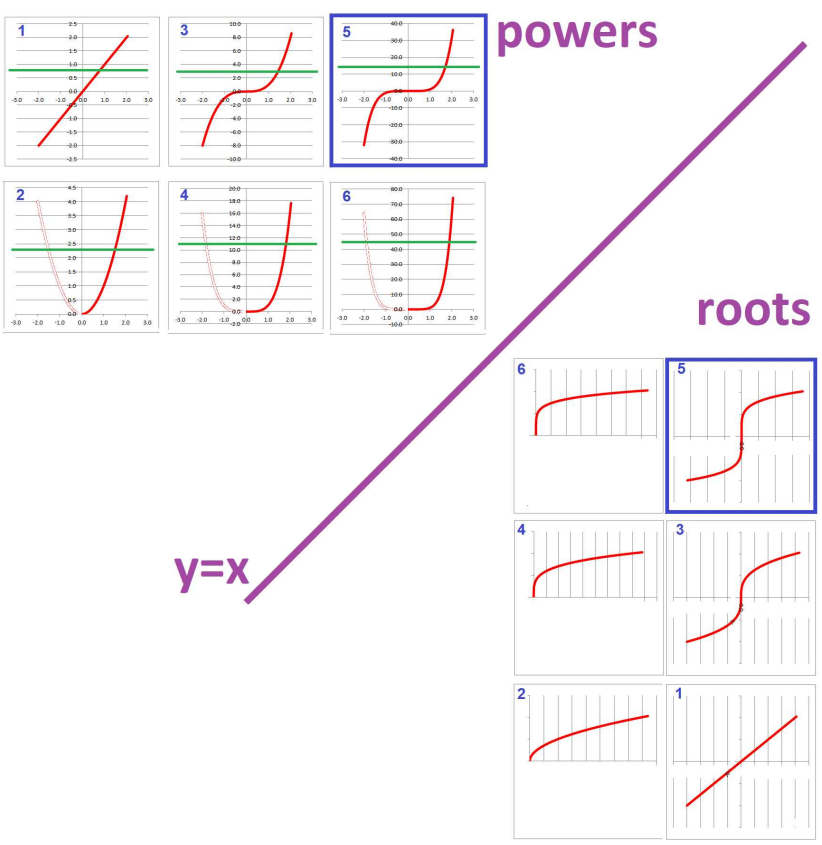
$$x,\ x^2,\ x^3,\ \ldots,$$

by applying some algebraic operations:



What if – in addition to these operations – we also consider the *inverses of the power functions*?

Our experience with $y = x^2$ tells us how to do this all at once. We flip the graphs of *all* power functions about the line $y = x$ at once:



Of course, the even degree powers aren't one-to-one! To make them invertible, we cut their domains, and codomains, to $[0, +\infty)$:

name of the function	formula			domain
The *square root*,	$y = \sqrt{x}$,	is the inverse of	$x = y^2$,	$x, y \geq 0$.
The *cubic root*,	$y = \sqrt[3]{x}$,	is the inverse of	$x = y^3$,	all x, y .
The *fourth degree root*,	$y = \sqrt[4]{x}$,	is the inverse of	$x = y^4$,	$x, y \geq 0$.
...				
The *nth degree root*,	$y = \sqrt[n]{x}$,	is the inverse of	$x = y^n$,	all x, y when n is add and $x, y \geq 0$ when n is even.
...				

The definition below is just another way to say the same.

Definition 4.7.2: nth root

For a real number x, the *nth root of* x is such a number y that $y^n = x$, denoted by

$$y = \sqrt[n]{x}$$

We require $x \geq 0$ when n is even.

In other words, y is a solution of the equation $y^n = x$. There is only one such solution!

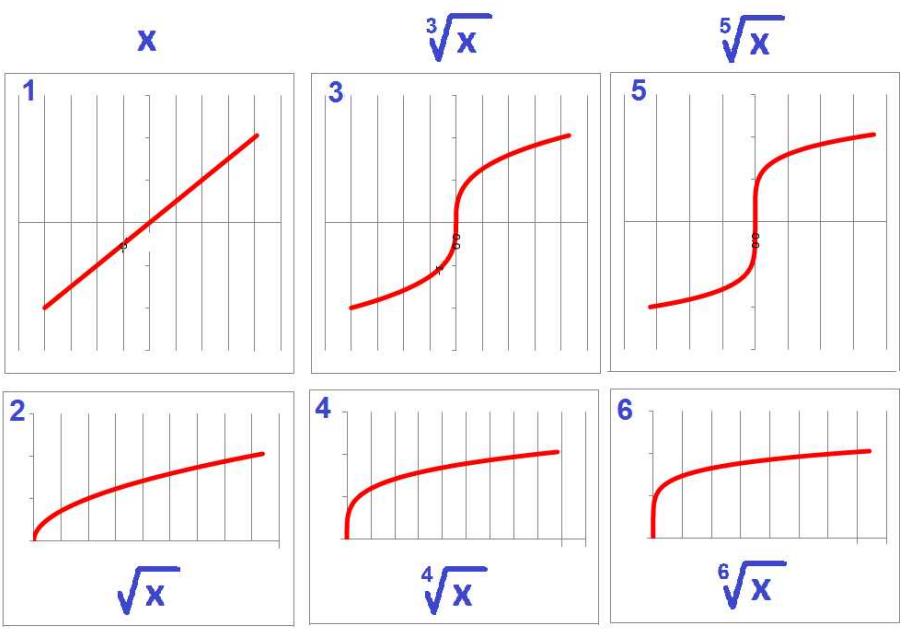
This is the way to look at these expressions as functions.

Roots

$$\sqrt[n]{\text{input}} = \text{output}$$

The symbol $\sqrt{\ }$ is called the "radical".

We imagine that we have solved all of these equations, for each value of x, and created a whole sequence of functions of x:



The illustration suggests the following:

Theorem 4.7.3: Odd Degree Roots

The roots of odd degrees

$$y = \sqrt[n]{x},\ n \text{ odd},$$

have the domain and range equal to $(-\infty, +\infty)$. *They are also strictly increasing, one-to-one, and odd.*

Theorem 4.7.4: Even Degree Roots

The roots of even degrees

$$y = \sqrt[n]{x},\ n \text{ even},$$

have the domain and range equal to $[0, +\infty)$. *They are also strictly increasing, one-to-one, and* ***not*** *even.*

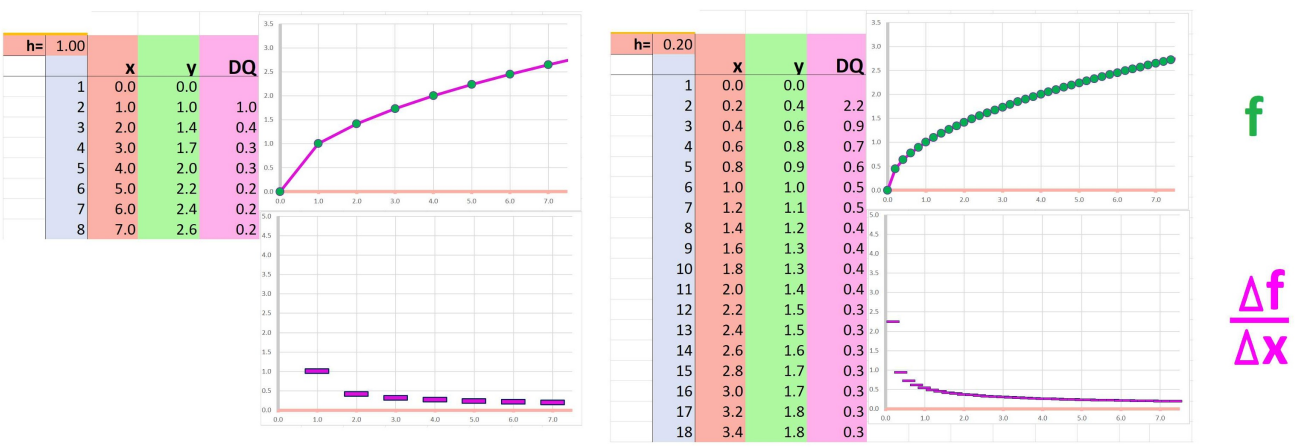
Because of the drastic step of cutting the domain, the symmetry of the even powers is lost...

Exercise 4.7.5

Prove the theorems.

We have classified the root functions according to these broad categories.

This is what the difference quotient of a root function looks like:



Exercise 4.7.6

Does the difference quotient look like any familiar function?

Now we turn to *algebra*.

Because the root functions are the inverses of the power functions, their algebraic properties are parallel. For example, we know that we can "distribute" powers (in other words, exponents, as discussed in Chapter 1) over multiplication:

$$(A \cdot B)^n = A^n \cdot B^n\,.$$

As it turns out, we can also "distribute" roots over multiplication, as follows.

Theorem 4.7.7: Product of Roots

For any integer n, *we have:*

$$\sqrt[n]{a \cdot b} = \sqrt[n]{a} \cdot \sqrt[n]{b}$$

provided $a, b \geq 0$ *whenever* n *is even.*

Proof.

We use the properties of the integer exponents (repeated multiplication) presented in Chapter 1. Suppose

$$A = \sqrt[n]{a}, \quad B = \sqrt[n]{b}\,.$$

Then:

$$a = A^n, \quad b = B^n\,.$$

Therefore, we have:

$$ab = A^n \cdot B^n = (AB)^n\,.$$

Then:

$$\sqrt[n]{ab} = AB\,.$$

For powers or roots, the expression *splits*.

Example 4.7.8: formulas are shortcuts

We use the rule as a shortcut, to expand:

$$\sqrt{20} = \sqrt{4 \cdot 5} = \sqrt{4} \cdot \sqrt{5} = 2\sqrt{5}\,.$$

and to contract:

$$\sqrt[5]{2}\sqrt[5]{2} = \sqrt[5]{2 \cdot 2} = \sqrt[5]{4}\,.$$

Exercise 4.7.9

Expand:

$$\sqrt[4]{400} = \,?$$

Exercise 4.7.10

Contract:

$$\sqrt[3]{2}\sqrt[2]{3} = \sqrt[6]{\,?\,}$$

Warning!

We can't "distribute" roots (nor powers) over addition:

$$\sqrt[n]{a+b} \neq \ldots$$

Exercise 4.7.11

Is there a way to simplify this: $(A+B)^n$?

Next, we know that the powers (exponents) are multiplied when composed:

$$(Q^n)^m = Q^{n \cdot m}\,.$$

The same happens with the degrees of roots.

Theorem 4.7.12: Composition of Roots

For any integers n and m, we have:

$$\sqrt[n]{\sqrt[m]{a}} = \sqrt[nm]{a}$$

provided $a \geq 0$ whenever n or m is even.

Proof.

We use the properties of the integer exponents (repeated multiplication) presented in Chapter 1. Suppose

$$A = \sqrt[m]{a}, \quad Q = \sqrt[n]{A}\,.$$

Then:

$$a = A^m, \quad A = Q^n\,.$$

Therefore, we have:

$$Q^{nm} = (Q^n)^m = (A)^m = a\,.$$

Then:

$$\sqrt[nm]{a} = Q\,.$$

For powers or roots, the composition is replaced with multiplication. The degrees m and n are presented in the opposite order to facilitate the proof (the latter from the former). We can watch them cancel below:

$$x \to \boxed{n\text{th power}} \to \boxed{m\text{th power}} \to \boxed{m\text{th root}} \to \boxed{n\text{th root}} \to y$$

Indeed, the two in the middle are inverses and undo each other:

$$x \to \boxed{n\text{th power}} \to \quad \to \quad \to \quad \to \quad \to \boxed{n\text{th root}} \to y$$

The two left are also inverses and also undo each other:

$$x \to \quad \to \quad \to \quad \to \quad \to \quad \to \quad \to \quad \to \quad \to y$$

Therefore, the composition of the first two functions is the inverse of the composition of the second two.

Example 4.7.13: formulas are shortcuts

We use the rule as a shortcut, to expand:

$$\sqrt[6]{64} = \sqrt[3\cdot 2]{64} = \sqrt[3]{\sqrt[2]{64}} = \sqrt[3]{8} = 2\,;$$

and to contract:

$$\sqrt[5]{\sqrt{2}} = \sqrt[5]{\sqrt[2]{2}} = \sqrt[5\cdot 2]{2} = \sqrt[10]{2}\,.$$

Exercise 4.7.14

Expand:

$$\sqrt[4]{100} = \,?$$

Exercise 4.7.15

Contract:

$$\sqrt[3]{\sqrt[3]{3}} = \,?$$

These theorems will be used later in this chapter to prove that every root can be seen as a (fractional) power, the reciprocal of the power the root came from.

Warning!

Though convenient, these two rules will be absorbed into the rules of exponents.

One can think of other rules:

$$\sqrt[m]{1} = 1\,, \quad \sqrt[m]{0} = 0\,, \text{ etc.}$$

Example 4.7.16: domain from flowchart

Consider the function

$$\sqrt{x-\sqrt{x}}\,.$$

This is its flowchart:

$$f:\ x\ \to\ \begin{array}{ccccccc} \nearrow & x & \to & x & \to & x & \searrow \\ & & & & & & & \boxed{-} \to z \\ \searrow & x & \to & \boxed{\sqrt{\ }} & \to & u & \nearrow \end{array}$$

Find the domain. Whatever is inside the square root can't be negative. So, we need to find all x for which we have:

$$x \geq 0 \quad \texttt{AND} \quad x-\sqrt{x} \geq 0\,.$$

Solve the latter assuming the former:

$$x \geq \sqrt{x} \implies x^2 \geq x \implies x \geq 1\,.$$

The domain is

$$D = \{x:\ x \geq 0\} \cap \{x:\ x \geq 1\} = [1,\infty)\,.$$

Exercise 4.7.17

Find the inverse of the function above and its domain.

With the help of the roots, we are now able to solve a lot more problems that involve powers. The nature of these problems is "inverse" to those we considered previously:

Models	**Direct problems**	**Inverse problems**
1. Square pool:	Q: What area if side 12 feet?	Q: How large side to have 300 sq feet?
$x \times x$ square with area	Substitute: $f(12) = 12^2$	Solve: $x^2 = 300\,.$
$f(x) = x^2$	$= 144\,.$	$\implies x = \sqrt{300} \approx 17$ feet.
2. Cubic tank:	Q: What volume if side 12 feet?	Q: How large side to have 300 cu feet?
$x \times x \times x$ cube with volume	Substitute: $f(12) = 12^3$	Solve: $x^3 = 300\,.$
$f(x) = x^3$	$= 1728\,.$	$\implies x = \sqrt[3]{300} \approx 6.69$ feet.

The solutions of the inverse problems are simply applications of the definition of the root!

Example 4.7.18: average rate of growth

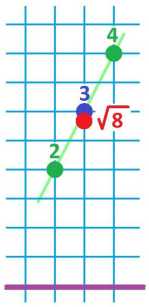
When we say that a population of bacteria grows at rate r per day, we just multiply it by r every day.

Suppose now that the population doubled yesterday and quadrupled today. What is its *average* rate of growth? It would be naive to answer that the rate has been $\dfrac{2+4}{2} = 3\,.$ The answer would suggest that the population has tripled every day! But tripled twice means that it has gone up by a factor of $3 \cdot 3 = 9$, not $2 \cdot 4 = 8$!

The correct answer is such a number x that when multiplied twice (squared) produce 8:

$$x \cdot x = 8 \implies x = \sqrt{8} \approx 2.83\,.$$

The difference is visible below:



Example 4.7.19: average return of investment

Suppose a mutual fund reports a 10% return last year and 5% for the year before. In order to evaluate its performance, we need to understand the meaning of the *average* return. It's not $\frac{10+5}{2} = 7.5\%$!

This return, when applied twice, should bring exactly the same as these two, i.e., the growth of $1.05 \cdot 1.10$. We, therefore, need such an x that

$$x \cdot x = 1.05 \cdot 1.10 = 1.155\,.$$

We solve:

$$x = \sqrt{1.155} \approx 1.0747\,.$$

So, the answer is "about 7.47%".

Example 4.7.20: average of two numbers

Suppose a quantity has grown from 1 to 9 over a period of time. Now, what was it half-way through this period? It depends! Consider these two possibilities:

- It has "grown by 8". Then half-way it has grown by this amount: $8/2 = 4$. Then, we have: $1 + 4 = 5$.
- It has "grown 9-fold". Then half-way it has grown by this factor: $\sqrt{9} = 3$. Then, we have: $1 \cdot 3 = 3$.

In summary, the question "what is the average of two numbers?" depends on the algebra they are involved in!

The difference is between an arithmetic progression and a geometric progression:

Question	Answer
How do you add *a in two steps*?	Add $\frac{a}{2}$ twice.
How do you multiply by *a in two steps*?	Multiply by $\sqrt{a}$ twice.

Exercise 4.7.21

Show that the latter is always smaller than the former.

That is why the root value always lies slightly below the line that connects the two points on the graph.

Analogy: repeated addition vs. repeated multiplication

	Multiplication:	**Exponentiation:**
	repeated addition	repeated multiplication
x is the average of a and b :	$a+b=x+x$ $\implies x=\dfrac{a+b}{2}$	$a\cdot b=x\cdot x$ $\implies x=\sqrt{a\cdot b}$
name:	*arithmetic mean*	*geometric mean*
general formula:	$\dfrac{1}{n}(a_1+a_2+\ldots+a_n)$	$\sqrt[n]{a_1\cdot a_2\cdot\ldots\cdot a_n}$

4.8. The exponential functions

Recall what we learned about the algebra of exponents in Chapter 1. Below is the summary of the *conventions* we have set up (for arbitrary $a,b>0$):

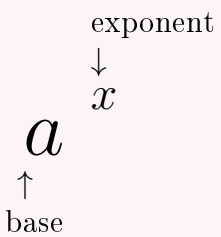
Analogy: repeated addition vs. repeated multiplication

	Multiplication:	**Exponentiation:**
$x=1,2,\ldots$	$\underbrace{a+a+a+\ldots+a}_{x\text{ times}}=a\cdot x$	$\underbrace{a\cdot a\cdot a\cdot\ldots\cdot a}_{x\text{ times}}=a^x$
$x=0$	$a\cdot 0=0$	$a^0=1$
$x=-1,-2,\ldots$	$ax=(-a)(-x)$	$a^x=\left(\dfrac{1}{a}\right)^{-x}=\dfrac{1}{a^{-x}}$
Rules: 1.	$a(x+y)=ax+ay$	$a^{x+y}=a^xa^y$
2.	$(a+b)x=ax+bx$	$(ab)^x=a^xb^x$
3.	$a(xy)=(ax)y$	$a^{xy}=(a^x)^y$

This is just a review of how the *power* – including the reciprocal powers – functions operate. However, our current interest is a different kind of function...

Recall the notation:

Base and exponent



The following is the new terminology we will use.

Definition 4.8.1: exponential function

The *exponential function* of base $a > 0$ is defined to be

$$f(x) = a^x$$

with the *domain* the set of all integers:

$$\mathbf{Z} = \{..., -3, -2, -1, 0, 1, 2, 3, ...\}\,.$$

Warning!

Unlike a power function, x^2, with a variable *base*, this is a function, 2^x, with a variable *exponent*.

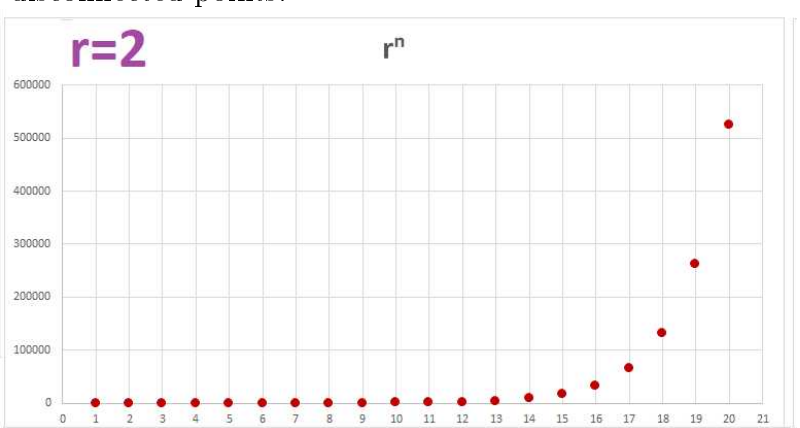
Unfortunately, the domain misses some of the numbers that interest us!

Example 4.8.2: bacteria still multiplying

Suppose we have a population of bacteria that doubles every day starting with $p_0 = 1$. We describe this as a geometric progression first (Chapter 1):

$$\underbrace{p_{n+1}}_{\text{population: at time } n+1} = 2 \cdot \underbrace{p_n}_{\text{at time } n} \implies p_n = 2^n\,.$$

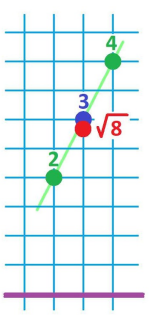
The graph is made of disconnected points:



Let's think of it as a function. It is given by the same formula, with x's still limited to the integers:

$$p(x) = 2^x\,.$$

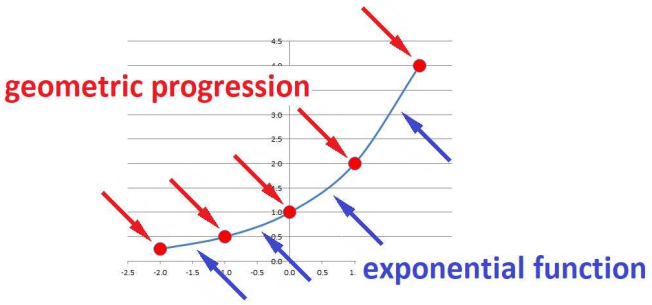
Now, what is the population in the middle of the first day? The function doesn't tell us (it is undefined at $x = .5$) but, according to the analysis in the last section, the answer should be $\sqrt{2}$. In other words, we choose the geometric mean over the arithmetic mean:



Does this mean that

$$p(1/2) = \sqrt{2} \quad \text{and} \quad 2^{1/2} = \sqrt{2}\,?$$

The domain of the exponential function $y = a^x$ with base $a > 0$ is all integers. What do we do about these gaps? Can, or should, we fill them, in a meaningful manner?



We figured out in Chapter 1 the meaning of "multiplication by a, zero times". We also figured out the meaning of "multiplication by a, a negative number of times". But what "multiplication by a one half times" can possibly mean?!

Example 4.8.3: compounded interest, revisited

Consider what happens to a \$1000 deposit with 10% annual interest, compounded yearly (Chapter 1):

$$\begin{aligned} \$1000, 1000 \cdot 1.10 &= 1000 + 10\% \text{ of } 1000 \\ \text{begin, after 1 year} &= \text{principal} + \text{interest} \\ &= 1000 + 1000 \cdot .10 \\ &= 1000(1 + 0.1) \\ &= 1000 \cdot 1.1\,. \end{aligned}$$

It's a geometric progression but this time we write is as a function; after x years we have:

$$f(x) = 1000 \cdot 1.1^x\,.$$

where x is a positive integer.

Now, what if I want to withdraw my money in the *middle* of the year? It would be fair to request from the bank for the interest to be compounded now. It would also be fair for the bank to want to do it in such a way that the annual return remains the same even if we compound twice. Accepting 5% interest would produce $1.05^2 = 1.1025$, or 10.25% annual interest.

Then, what should be this semi-annual interest rate? Suppose the amount has grown by a proportion, r, so that, if applied again, it will give me the same ten percent growth! In other words, we have:

$$f(.5) = 1000 \cdot r \quad \text{and} \quad r \cdot r = 1.1\,.$$

Therefore, according to the definition of the square root, we have

$$r = \sqrt{1.1} \approx 1.0488\,,$$

or about 4.9 percent.

Exercise 4.8.4

What if you want to compound quarterly?

Exercise 4.8.5

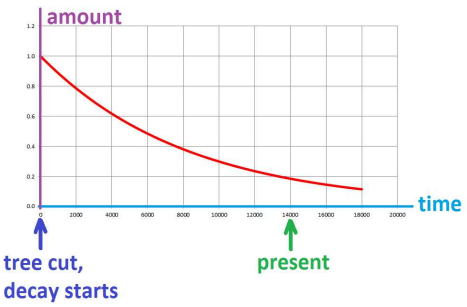
If your bank promises to pay you 1% for the first year and 2% for the second, what *single* (annual) interest can it offer in order to pay you as much over the next two years? Explain the meaning of the *average* interest rate.

Example 4.8.6: radioactive decay and radiocarbon dating, revisited

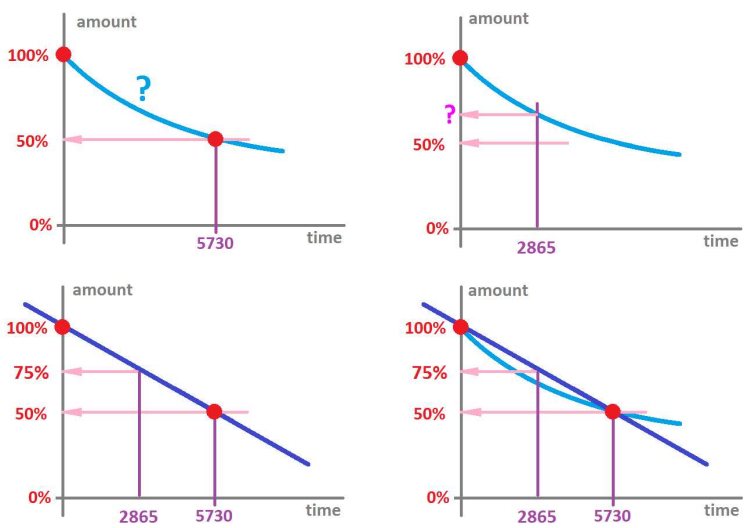
Recall an example from Chapter 1. The radioactive carbon loses half of its mass over a certain period of time called the *half-life* of the element. It's a geometric progression again:

$$a_{n+1} = a_n \cdot \frac{1}{2}\,.$$

Unfortunately, n is not the number of years but the number of half-lives! For example, the percentage of this element, ^{14}C, left plotted below against time may look like this:



However, we only know *two* points on the graph! Suppose the half-life is 5730 years (i.e., the time it takes to go from 100% to 50%). The model measures time in multiples of the half-life, 5730 years, and any period shorter than that will require a new insight. Before we even try to date a parchment with, say, 75% of ^{14}C left, let's try to ask a simpler question: How much is left after $5730/2 = 2865$ years?



The linear (or arithmetic) answer is that the loss is half of the half. It is, therefore, a quarter, and what's left is 75%! The real number is lower:

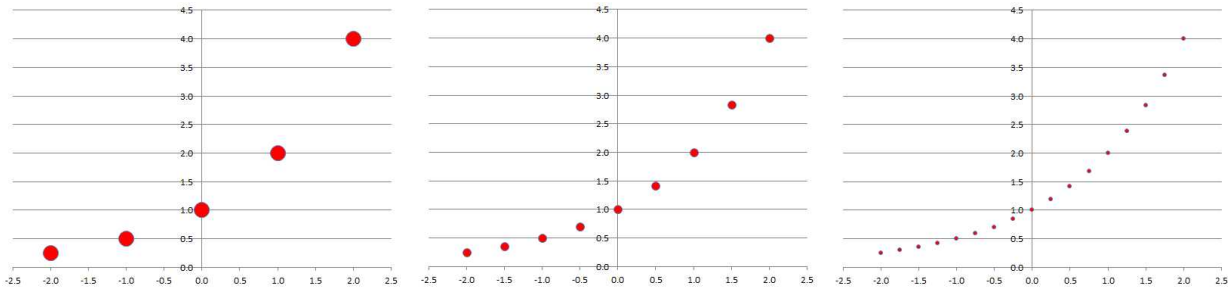
$$\sqrt{\frac{1}{2}} \approx .707\,.$$

Example 4.8.7: square root as average

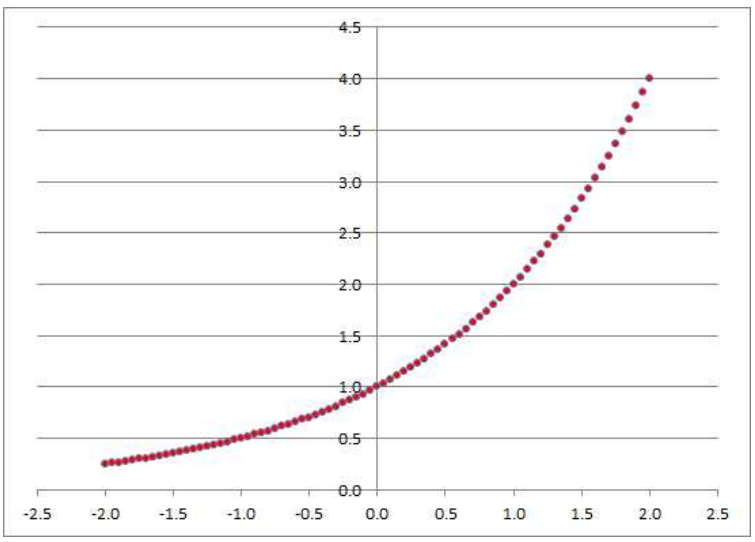
For any two numbers a and b, we know the value of the exponent for the number half-way between them. For example, we have the following:

$$x = \frac{a+b}{2} \implies 2^x = \sqrt{2^a \cdot 2^b}\,.$$

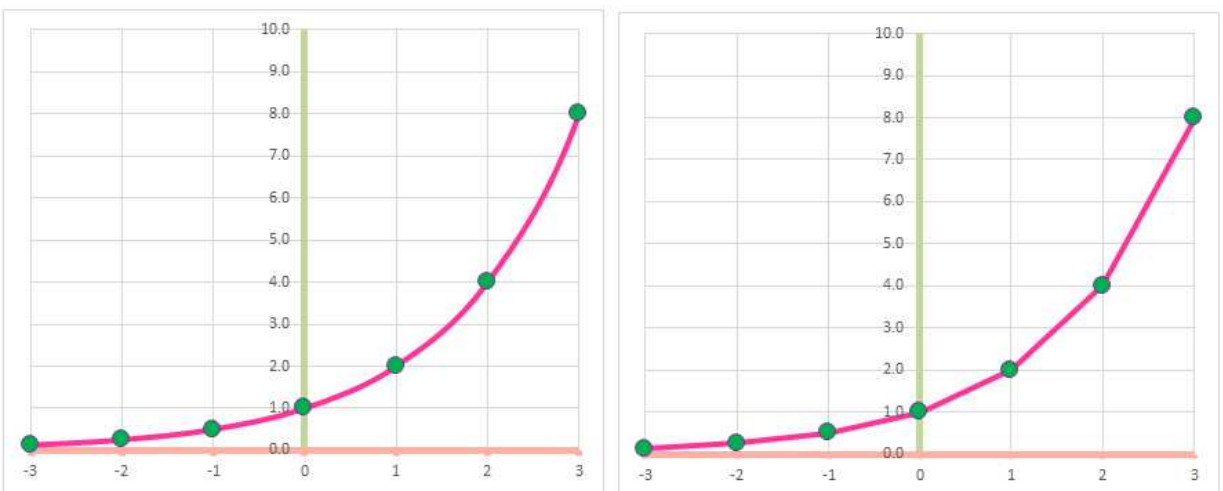
What if we continue to produce more and more values of this function by dividing the intervals in half? The new value will always lie slightly below the line that connects the two points on the graph:



One can also imagine that the x-axis, as the domain, is becoming denser and denser covered. These initially loose points seem to start to form a curve:



The function that we have designed via the geometric mean is compared to the one that uses the arithmetic mean:



The graph of our function (left) lies below any chord that connects two points on the graph (right). Such a function is called "concave up" (Chapter 2DC-3).

Example 4.8.8: nth root

What if we divide the interval into 3 parts instead of 2? What is the value of 2^x if $x = 1/3$? We follow the same idea:

$$2^{\frac{1}{3}} = \sqrt[3]{2}\,.$$

Furthermore, we can divide into smaller and smaller parts; then:

$$2^{\frac{1}{n}} = \sqrt[n]{2}\,.$$

As a new convention, we define the reciprocal exponent, as follows. For any positive integer n, we set the

$\frac{1}{n}$*th power* of $a > 0$, or *a taken to the power* $1/n$, to be the following:

$$a^{1/n} = \sqrt[n]{a}$$

Exercise 4.8.9

Show that rules 2 and 3 in the above table above still hold:

$$a^x b^x = (ab)^x, \ a^{xy} = (a^x)^y \, .$$

Having the reciprocals as exponents isn't enough (what is $2^{10.5}$?) but it is a step in the right direction. Our new convention is to include all *rational numbers*, i.e., fractions of integers, with the ultimate goal to have the domain to be $(-\infty, \infty)$.

Definition 4.8.10: rational exponent

If $x = \frac{m}{n}$, where m, n are integers with $n > 0$, then we set the *xth power* of $a > 0$ to be the following:

$$a^{\frac{m}{n}} = \sqrt[n]{a^m}$$

It also reads "a to the power of x".

Warning!

Unlike geometric progressions, an exponential function can only have a positive base, $a > 0$; no $(-1)^x$ anymore!

We now need to show that the rules still apply to *all* rational numbers.

Theorem 4.8.11: Addition-Multiplication Rule of Exponents

For every real $a > 0$ and every rational x and y, we have:

$$a^{x+y} = a^x a^y$$

Exercise 4.8.12

Prove the formula.

Theorem 4.8.13: Distributive Rule of Exponents

For every real $a, b > 0$ and every rational x, we have:

$$a^x b^x = (ab)^x$$

Proof.

Suppose $x = \frac{1}{n}$, where n is a positive integer. Then:

$$a^x b^x = a^{\frac{1}{n}} b^{\frac{1}{n}} = \sqrt[n]{a}\sqrt[n]{b} = \sqrt[n]{ab} = (ab)^{\frac{1}{n}} = (ab)^x \, ,$$

according to the *Product of Roots* theorem in the last section. We have the formula proven for the reciprocals.

Next, suppose $x = \dfrac{m}{n}$, where m is an integer and n is a positive integer. Then:

$$a^x b^x = a^{\frac{m}{n}} b^{\frac{m}{n}} = \sqrt[n]{a^m}\sqrt[n]{b^m} = \sqrt[n]{a^m b^m} = \left((ab)^m\right)^{\frac{1}{n}} = (ab)^{\frac{m}{n}} = (ab)^x\,,$$

according to rule for the reciprocals.

Theorem 4.8.14: Multiplication-Exponentiation Rule of Exponents

For every real $a > 0$ and every rational x and y, we have:

$$a^{xy} = (a^x)^y$$

Proof.

Suppose $x = \dfrac{1}{n}$ and $y = \dfrac{1}{m}$, where n, m are positive integers. Then:

$$(a^x)^y = \left(a^{\frac{1}{n}}\right)^{\frac{1}{m}} = \sqrt[m]{\sqrt[n]{a}} = \sqrt[nm]{a} = a^{\frac{1}{nm}} = a^{\frac{1}{n}\frac{1}{m}} = a^{xy}\,,$$

according to the *Composition of Roots* theorem in the last section. We have the formula proven for the reciprocals.

Exercise 4.8.15

Provide the missing parts of the proof.

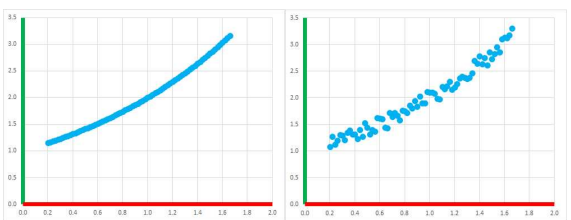
This is the idea of our formula:

Rational exponent

$$a^{\frac{m}{n}} = \sqrt[n]{a^m}$$

Is the graph of this new function a "complete" curve such that of $y = x^2$?

We have seen that, as these fractions get larger and larger denominators, the domain is becoming denser and denser and, eventually, the graph becomes a curve! After all, the domain contains all rational numbers $\mathbf{Q}$ and the rational numbers are so *dense* that any interval, no matter how small, will contain infinitely many of them (after all, if p and q are rational, then so is $(p+q)/2$). Therefore, there are *no gaps* in the graph! However, its points remain *disconnected* from each other; a little blow and the curve falls apart:



The reason is that the points on the graph, unlike that of $y = x^2$, with *irrational* x-coordinates are missing:

$$x = \sqrt{2},\ \sqrt{3},\ \pi\,.$$

And the irrational numbers are, too, so *dense* that any interval, no matter how small, will contain infinitely many of them. So, even though there are no gaps in the graph, there are invisible "cuts" everywhere! We can define the function for the irrational exponents by "approximating" them with rational numbers (to be presented in Chapter 2DC-1). For example, the sequence 2^3, $2^{3.1}$, $2^{3.14}$, ... will approximate 2^π. The three algebraic rules are still to be obeyed.

Warning!

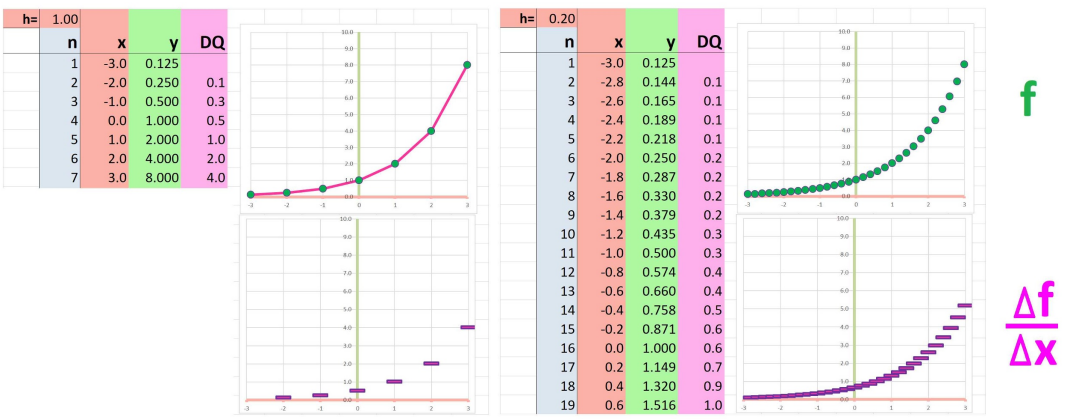
Even though this function has been built from the power functions and the roots, it's very different from those.

We start treating the exponential function as if it has been already fully constructed.

Exercise 4.8.16

Describe such a function as a transformation.

Below we plot the difference quotient $\dfrac{\Delta f}{\Delta x}$, i.e., the sampled slopes, of $f(x) = 2^x$:



h=	1.00			
	n	x	y	DQ
	1	-3.0	0.125	
	2	-2.0	0.250	0.1
	3	-1.0	0.500	0.3
	4	0.0	1.000	0.5
	5	1.0	2.000	1.0
	6	2.0	4.000	2.0
	7	3.0	8.000	4.0

h=	0.20			
	n	x	y	DQ
	1	-3.0	0.125	
	2	-2.8	0.144	0.1
	3	-2.6	0.165	0.1
	4	-2.4	0.189	0.1
	5	-2.2	0.218	0.1
	6	-2.0	0.250	0.2
	7	-1.8	0.287	0.2
	8	-1.6	0.330	0.2
	9	-1.4	0.379	0.2
	10	-1.2	0.435	0.3
	11	-1.0	0.500	0.3
	12	-0.8	0.574	0.4
	13	-0.6	0.660	0.4
	14	-0.4	0.758	0.5
	15	-0.2	0.871	0.6
	16	0.0	1.000	0.6
	17	0.2	1.149	0.7
	18	0.4	1.320	0.9
	19	0.6	1.516	1.0

In contrast to all other functions we have seen, the difference quotient appears to exhibit a behavior similar to the original function! We will show (Chapter 2DC-3) that the rate of growth of an exponential function grows exponentially.

Exercise 4.8.17

Plot the difference quotient of the logistic function (i.e., restricted growth):

$$f(x) = \frac{1}{1+2^{-x}}\,.$$

What pattern do you see?

Below is the summary of our *conventions* ($a > 0$ real, m, n positive integers):

Analogy: repeated addition vs. repeated multiplication

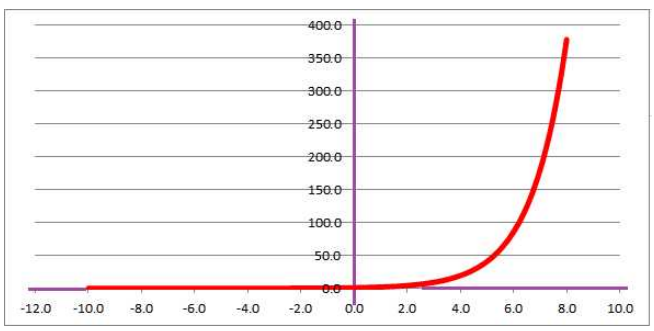
		Multiplication:	**Exponentiation:**
1.	$x = 1, 2, ...$	$ax = \underbrace{a + a + a + ... + a}_{x \text{ times}}$	$a^x = \underbrace{a \cdot a \cdot a \cdot ... \cdot a}_{x \text{ times}}$
2.	$x = 0$	$a \cdot 0 = 0$	$a^0 = 1$
3.	$x = -1, -2, ...$	$ax = (-a)(-x) = -a(-x)$	$a^x = \left(\frac{1}{a}\right)^{-x} = \frac{1}{a^{-x}}$
4.	$x = \frac{m}{n}$	$ax = \frac{am}{n} = \frac{a}{m}n$	$a^x = \sqrt[n]{a^m} = \left(\sqrt[n]{a}\right)^m$
5.	x real		Chapter 2DC-1

As we progress from the natural numbers to the real, we also progress from sequences to functions in our analogy:

Analogy: repeated addition vs. repeated multiplication

Multiplication:	**Exponentiation:**
arithmetic progression, $a_{n+1} = a_n + a$	geometric progression, $a_{n+1} = a_n \cdot a$
linear function, $f(x) = mx + b$	exponential function, $f(x) = a^x$

Let's collect some facts about this function visible in the graph of $y = 2^x$:



Theorem 4.8.18: Facts About Exponential Function

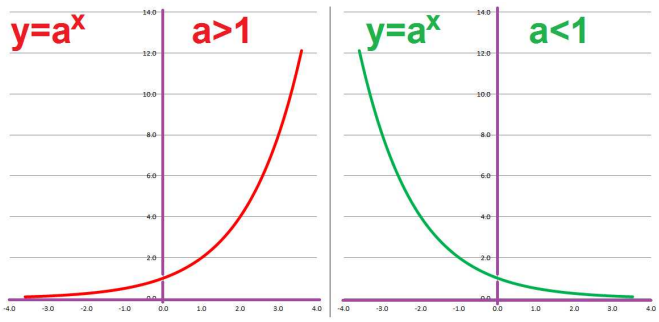
The exponential function $y = a^x$ with base $a > 0$ satisfies the following:

1. *The domain is $(-\infty, +\infty)$.*
2. *The range is $(0, \infty)$.*
3. *The y-intercept is $(0, 1)$.*
4. *There are no x-intercepts.*

Exercise 4.8.19

Prove parts 2-4.

There are two very different kinds of exponential functions however:



The following is an extension of the theorem about the monotonicity of the geometric progression.

Theorem 4.8.20: Monotonicity of Exponent

The exponential function with base $a > 0$ satisfies the following:

$$y = a^x \text{ is strictly increasing if } a > 1\,,$$
$$y = a^x \text{ is strictly decreasing if } a < 1\,.$$

Proof.

Suppose $a > 1$. The theorem has already been proven for the case of integer x. Now, suppose x and y are rational numbers and $x < y$. Suppose

- $x = n/m$ for two integers m, n, and
- $y = p/q$ for two integers p, q.

Then:

$$n/m < p/q \text{ , or } nq < mp\,.$$

Then:

$$a^{nq} < a^{mp}\,,$$

because the exponential function is increasing for integer inputs. Therefore:

$$\sqrt[q]{a^{nq}} < \sqrt[q]{a^{mp}}\,,$$

because the roots are increasing functions. Then:

$$a^n < a^{mp/q}\,.$$

Similarly, we obtain:

$$\sqrt[m]{a^n} < \sqrt[m]{a^{mp/q}} \text{ , or } a^{n/m} < a^{p/q}\,.$$

These two behaviors are called, respectively,

- "exponential growth" (population growth, compounded interest), and
- "exponential decay" (population decline, radioactive decay, warming and cooling).

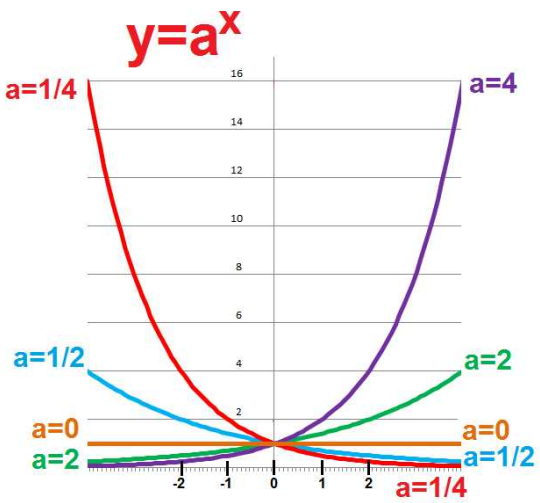
Corollary 4.8.21: Exponential Function Is Invertible

The exponential function is one-to-one and, therefore, invertible as a function

$$f : (-\infty, +\infty) \to (0, +\infty)\,.$$

The inverse of this function is discussed in the next section.

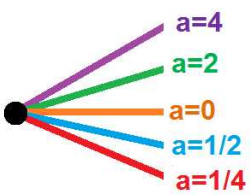
Within the two classes, the exponential functions still vary according to their bases: the larger a is, the faster a^x grows (Chapter 2DC-3):



Exercise 4.8.22

What is the union of all these graphs?

The constant function separates the exponential growth functions from the exponential decay functions:



there are infinitely many exponential functions! But there may be more patterns to discover about these functions! Let's look at them as a whole. We will discover that they are all interconnected!

Example 4.8.23: horizontal flip of exponential function

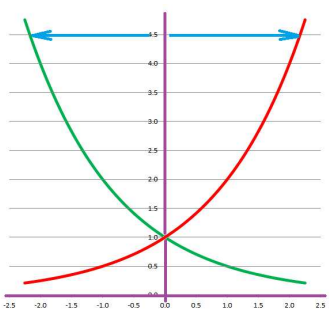
Let's plot

$$y = \left(\frac{1}{2}\right)^x .$$

We save time by making an observation:

$$\left(\frac{1}{2}\right)^x = 2^{-x} .$$

The point is that the graph of our function (green) is that of the familiar $y = 2^x$ (red) flipped about the y-axis ((x, y) becomes $(-x, y)$):



We have discovered the following:

- ▶ The graph of the exponential function of base a, $y = a^x$, flipped about the y-axis produces the graph of the exponential function of the reciprocal base $1/a$, $y = (1/a)^x$.

Conversely, all exponential decay functions can be acquired from the exponential growth function in this

manner. And, all exponential growth functions can be acquired from the exponential decay function!

Exercise 4.8.24

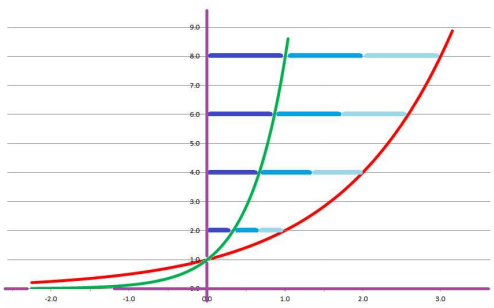
Prove these statements.

Example 4.8.25: horizontal stretch/shrink of exponential function

Consider:

$$y = 2^{3x}\,.$$

Once we have plotted, point by point, its graph, we realize that it looks very similar to the exponential growth functions we have seen:



A bit of algebra – the *Multiplication-Exponentiation Rule* – reveals the connection:

$$2^{3x} = (2^3)^x = 8^x\,.$$

It *is* an exponential function (base 8)! Therefore, the graph of $y = 8^x$ is that of $y = 2^x$ shrunk horizontally by a factor of 3.

We have discovered the following:

► The graph of the exponential function of base a, $y = a^x$, shrunk horizontally by a factor of k produces the graph of the exponential function of the base a^k, $y = (a^k)^x$.

Exercise 4.8.26

Prove this statement. What is the converse?

We will prove in the next section the converse: We can get all the exponential growth functions by stretching one of them.

Exercise 4.8.27

What about the *vertical* stretch?

To summarize the analysis, we say:

► Horizontally stretching and flipping the graph of an exponential function produces the graph of another exponential function.

Conversely:

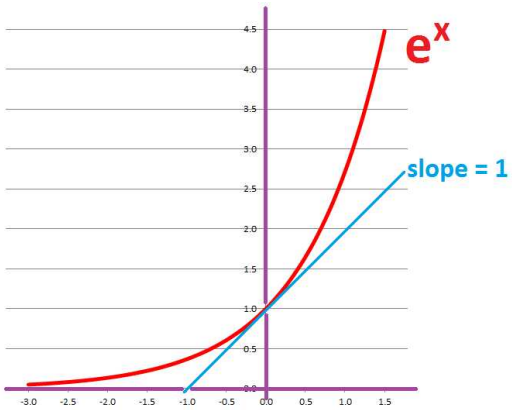
► We can get the graphs all exponential functions by stretching and flipping one of them horizontally.

So, just as we need only one quadratic polynomial, $y = x^2$, as a *template* for the rest, we really only need one exponential function!

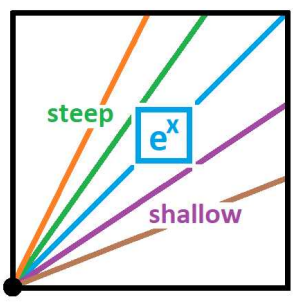
Which one is the *one*? There is a special number, the *Euler's number*:

$$e \approx 2.71828\,.$$

It produces e^x, the *natural base exponent* (Chapter 2DC-1). What makes it special is this property: its graph crosses the y-axis at 45 degrees (Chapter 2DC-3):



It also cuts, in a sense, the exponential functions into two halves: the "steep" ones and the "shallow" ones:



Example 4.8.28: population growth, revisited

Let's consider the population growth example again (Chapter 1). Suppose there are 50 babies born per 10,000 of population. The proportion is

$$\frac{50}{10000} = 0.005\,.$$

Therefore, after x years, the population is predicted to become

$$1,000,000 \cdot 1.005^x\,.$$

Now, how long will it take for the population to *double*? We can use trial and error, but generally, we have to solve the following equation:

$$1,000,000 \cdot 1.005^x = 2,000,000\,.$$

We will, therefore, need the *inverse* of the function we just built, presented in the next section.

Below is the summary of the rules of exponents that will reappear many times.

Theorem 4.8.29: Rules of Exponents

The identities below are satisfied for each real $a > 0$ and every real x and y:

1. $a^{x+y} = a^x a^y$
2. $a^x b^x = (ab)^x$
3. $a^{xy} = (a^x)^y$

4.9. The logarithmic functions

Example 4.9.1: bacteria multiplying, continued

Bacteria double every day. How many times do they have to double in order to quadruple? Two.

How long does it take to increase 8-fold? Let's think. We start with the function that describes this process and set it equal to 8:

$$2^x = 8\,,$$

where x is the time in days. The answer is, of course, 3.

But what if it is 7-fold instead? Or, how many times do you double in order to triple? To answer these well-justified questions, we will need the *inverse* of the exponential function!

Recall that for the inverse to exist the function has to be one-to-one and onto. The exponential function is one-to-one (unlike x^2) and, therefore, we don't need to do anything with its *domain*. However, it isn't onto as it can only take positive values. The move that resolves the issue is to choose the *codomain* of a new function to be the range of the old function! Instead of

$$a^x : \mathbf{R} \to \mathbf{R}\,,$$

we consider

$$a^x : \mathbf{R} \to (0, +\infty)\,.$$

The following definition is then justified.

Definition 4.9.2: logarithm

For any real $a > 0$, the *logarithm* base a is the inverse of the exponential function base a, i.e., $y = a^x$, with codomain $(0, +\infty)$. The function is denoted by

$$x = \log_a y$$

Example 4.9.3: population growth, revisited

Armed with nothing but this definition – and a calculator – we can solve the problem in the last section. If the population is predicted to be after x years:

$$1,000,000 \cdot 1.005^x\,,$$

how long will it take for the population to *double*? Solving the equation:

$$1,000,000 \cdot 1.005^x = 2,000,000\,,$$

produces:

$$1.005^x = 2 \implies x = \log_{1.005} 2 \approx 139 \text{ years}.$$

Exercise 4.9.4

Use the definition and a calculator to find how long it takes to triple your money if your interest is 1.5%.

The link between the two functions inverse of each other will remain unbroken:

$$\mathbf{a}^{\mathbf{x}} = \mathbf{y} \iff \log_{\mathbf{a}} \mathbf{y} = \mathbf{x}$$

Example 4.9.5: definition of log

Every computed exponent generates a computed logarithm:

$$2^3 = 8 \iff \log_2 8 = 3\,.$$
$$10^2 = 100 \iff \log_{10} 100 = 2\,.$$

The bases remain unchanged but the inputs become outputs and vice versa.

Exercise 4.9.6

Compute: (a) $\log_2 1024$, (b) $\log_2 .25$, (c) $\log_1 1$.

Just as with the exponential function, in the notation for the logarithm *the base is placed below the input* of the function:

$$\text{base}^x \quad \text{and} \quad \log_{\text{base}} y$$

The input is a superscript of the base in the former case and the base is a subscript of the input in the latter case.

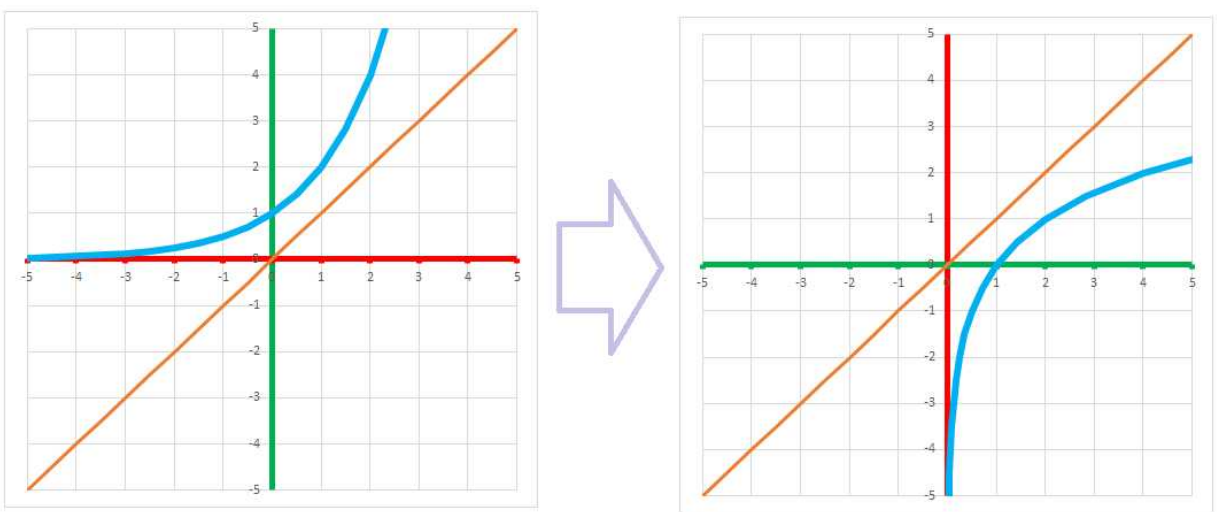
Warning!

Following our conventions about functions, we can alternatively write:

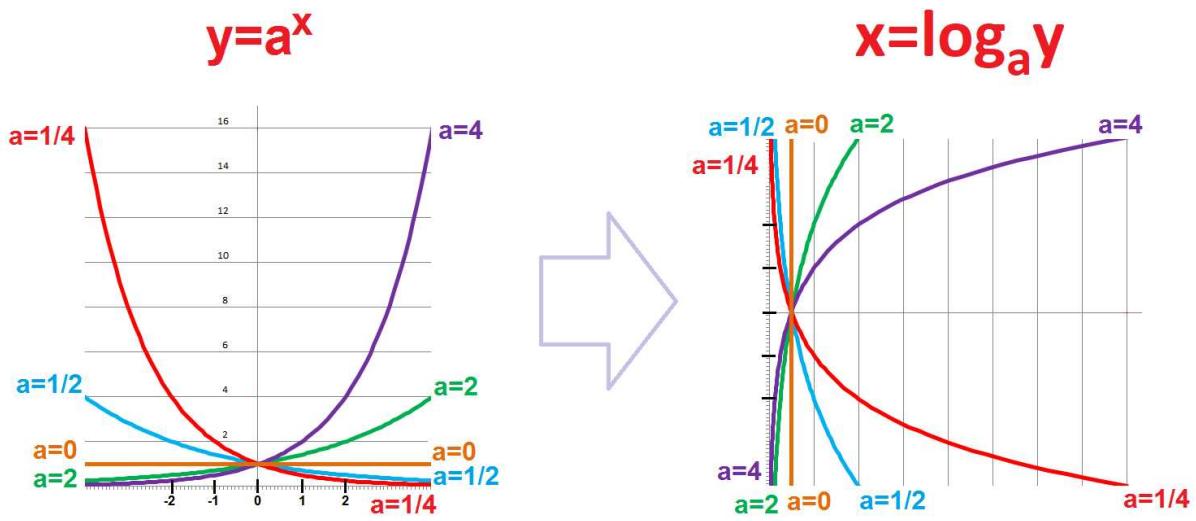
$$a^{(x)} \quad \text{and} \quad \log_a(x)$$

We have acquired, in one stroke, infinitely many new functions. Now the *graphs*.

This is how we get the graph of the logarithm with a flip about the diagonal $y = x$:



In fact, we can have *all* of the graphs at once:



From the basic features of the graph of the exponential function, we derive those for the logarithm; we just interchange x and y:

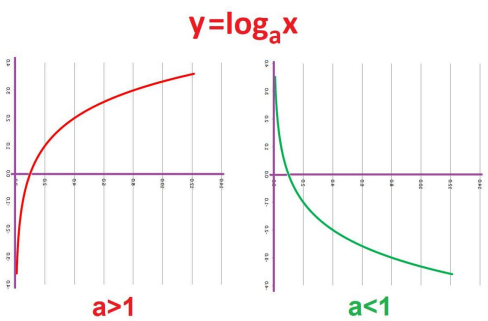
$y = a^x$	$y = \log_a x$
The domain is $(-\infty, +\infty)$.	The range is $(-\infty, +\infty)$.
The range is $(0, \infty)$.	The domain is $(0, \infty)$.
The y-intercept is $(0, 1)$.	The x-intercept is $(0, 1)$.
There are no x-intercepts.	There are no y-intercepts.

They are also all one-to-one.

Exercise 4.9.7

Do you see any asymptotes in the graphs?

Just as with the exponential function, there are two very different kinds of logarithms:



From the *Monotonicity of Exponent* theorem in the last section we derive the following about the logarithm.

Theorem 4.9.8: Monotonicity of Logarithm

The logarithm function with base $a > 0$ satisfies the following:

$y = \log_a x$ *is strictly increasing if* $a > 1$.
$y = \log_a x$ *is strictly decreasing if* $a < 1$.

In other words, the statement "x and y increase together" remains true if we interchange x and y.

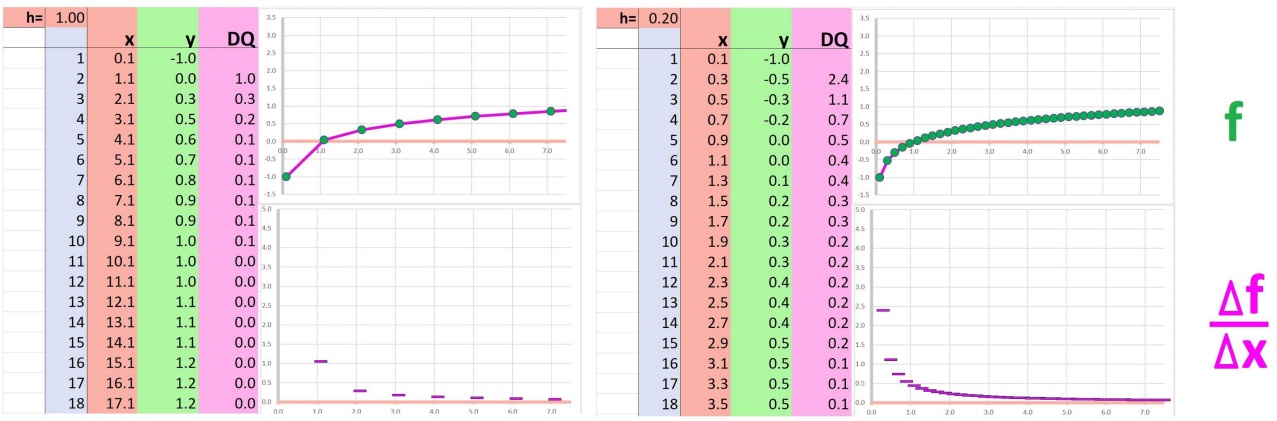
Exercise 4.9.9

Prove the theorem.

Exercise 4.9.10

Describe such a function as a transformation.

This is the difference quotient of the logarithm:



Exercise 4.9.11

Does it look familiar?

Now some *algebra*...

The logarithm base a of y is the number that can be defined *verbally* as follows:

> $\log_a y$ is the power to which you have to raise a to get y.

Therefore, we can say that these two below are the *same* statement written in two different ways:

$$a^x = y \quad \text{vs.} \quad \log_a y = x\,.$$

After all, the *relation* between x and y (Chapter 2) remains the same.

Example 4.9.12: logarithms come from prior exponents

For a fixed base, exponents and logarithms come in pairs:

$a^x = y$	means	$\log_a y = x\,.$
$2^3 = 8$	means	$\log_2 8 = 3\,.$
$10^2 = 100$	means	$\log_{10} 100 = 2\,.$
$2^{-1} = \frac{1}{2}$	means	$\log_2 \frac{1}{2} = -1\,.$
$3^{1/2} = \sqrt{3}$	means	$\log_3 \sqrt{3} = \frac{1}{2}\,.$

In a sense, every logarithmic expression used to be an exponential expression. This is why, initially, computing logarithms means *remembering* exponents computed in the past, just like with the roots earlier in this chapter:

- What is $\sqrt[4]{16}$? I seem to recall computing the powers of 2 and producing 16 after 4 repetitions. So, $\sqrt[4]{16} = 2$.
- What is $\log_3 27$? I seem to recall computing the powers of 3 and producing 27 in 3 steps. So, $\log_3 27 = 3$.

In more complex situations, we have to work our way to a representation of the input of the logarithm as a power of its base.

- What is $\log_{10} .01$? I need to represent .01 as a power of 10. Here it is: $.01 = \frac{1}{100} = \frac{1}{10^2} = 10^{-2}$. So, $\log_{10} .01 = -2$.

- What is $\log_4 2$? I need to represent 2 as a power of 4. Here it is: $2 = \sqrt{4} = 4^{1/2}$. So, $\log_4 2 = \dfrac{1}{2}$.

This, however, is hardly a *method*.

- What is $\log_4 5$? I need to represent 5 as a power of 4. There seems to be no easy way to do it. I know; here it is: $5 = 4^{\log_4 5}$! So, $\log_4 5 = \log_4 5$.

This "answer" is just as acceptable as the ones before; we are just unable to *simplify* this expression.

Warning!

There is no "formula" for the logarithm, just as there is no "formula" for the square root.

The fact that the exponential function and the logarithm – of the same base – are inverses of each other means that they *undo* each other when composed, in either order. For the pair

$$y = a^x \quad \text{and} \quad x = \log_a y\,,$$

we have these two rules.

Theorem 4.9.13: Cancellation Laws of Logarithms

Suppose $a > 0$. Then for any real x and any $y > 0$, we have:

$$\begin{aligned} a^{\log_a y} &= y \\ \log_a a^x &= x \end{aligned}$$

In other words, the outputs of these compositions are the same as the inputs:

$$x \;\to\; \boxed{a^x = y} \;\to\; y \;\to\; \boxed{\log_a y = x} \;\to\; x$$

and

$$y \;\to\; \boxed{\log_a y = x} \;\to\; x \;\to\; \boxed{a^x = y} \;\to y$$

These two flowcharts can be seen in the above formulas, as follows:

1. x
2. $\log_a(\ x\ )$
3. $a^{\ }$

1. y
2. $\log_a(\ y\ )$
3. $(\log_a(\ y\ ))$
4. $a^{\ }$

Warning!

The exponent and the logarithm cancel each other only if they are of the *same base*.

We use these formulas to solve *exponential equations*.

Example 4.9.14: solving exp equations with log

Solve the following for x:

$$2^{x-5} = 3\,.$$

To "kill" the exponent and get to x, apply its *inverse*, the logarithm of the same base, to both sides of

the equation:

$$\log_2(2^{x-5}) = \log_2(3)\,.$$

We are after this composition:

$$\log_2(2^{x-5}) = \log_2 3\,.$$

It now disappears after the second cancellation law:

$$\begin{array}{ll} x-5 & = \log_2 3 \qquad \Longrightarrow \\ x & = \log_2 3 + 5\,. \end{array}$$

Exercise 4.9.15

Solve the equation below.

$$3^{x^2+1} = 2\,.$$

Make up your own equation and solve it. Repeat.

Exercise 4.9.16

Suppose function f performs the operation "take the logarithm of", and function g performs "take the square root of". (a) Find the formulas for the two possible compositions. (b) Find their domains.

Example 4.9.17: solving log equations with exp

Solve the following for x:

$$\log_2(x-5) = 3\,.$$

To "kill" the logarithm and get to x, apply its *inverse*, the exponential function of the same base, to both sides of the equation:

$$2^{(\log_2(x-5))} = 2^{(3)}\,.$$

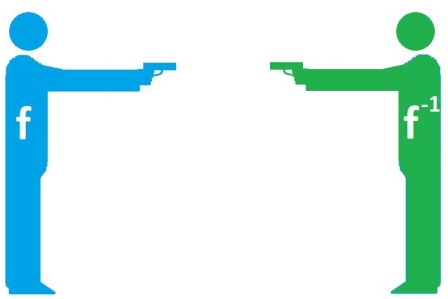
We are after this composition:

$$2^{\log_2(x-5)} = 8\,.$$

It now disappears after the first cancellation law:

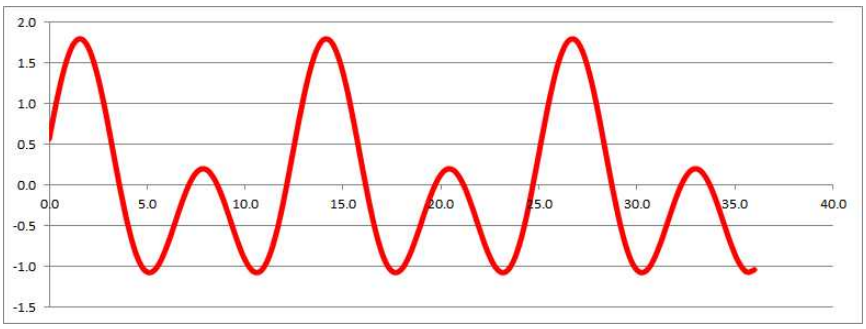
$$\begin{array}{ll} x-5 & = 8 \qquad \Longrightarrow \\ x & = 13\,. \end{array}$$

This kind of interaction between a function and its inverse is common:



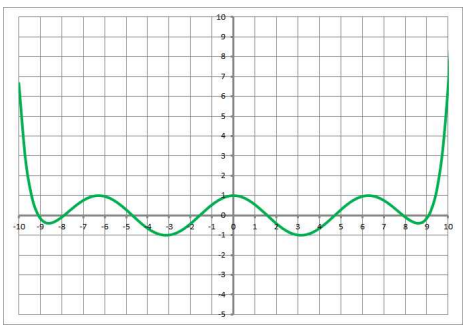
4.10. The trigonometric functions

One encounters numerous examples of *periodic phenomena*. The simplest case is that of a quantity that changes but then comes back to change again in the same manner:



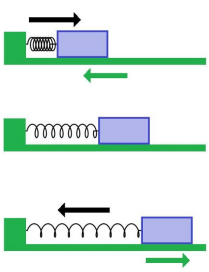
We call such functions *periodic*.

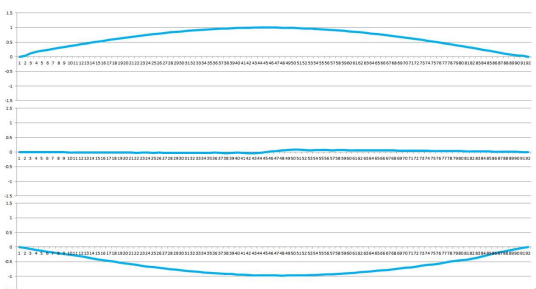
However, none of main classes of functions we have so far introduced exhibit this behavior. For example, a well-chosen polynomial can mimic periodicity but eventually will have to run away to infinity:

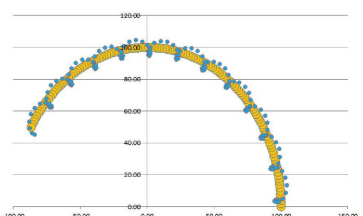


Example 4.10.1: periodic behavior

The simplest periodic behavior is oscillation of an object on a spring or a string of a musical instrument or an orbiting planet:

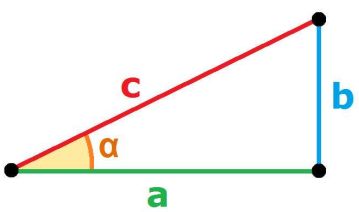






We introduce a new class of functions.

The trigonometric functions initially come from *plane geometry*:



Definition 4.10.2: sine, cosine, and tangent

Suppose we have a right triangle with sides a, b, c, with c the longest one facing the right angle. If α is the angle adjacent to side a, then we define the *cosine* and the *sine* of this angle as follows:

$$\begin{aligned} \cos\alpha &= \frac{a}{c} \\ \sin\alpha &= \frac{b}{c} \end{aligned}$$

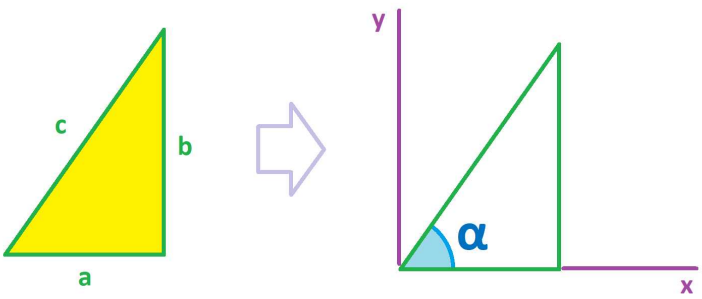
We also define its *tangent*:

$$\tan\alpha = \frac{b}{a} = \frac{\sin\alpha}{\cos\alpha}$$

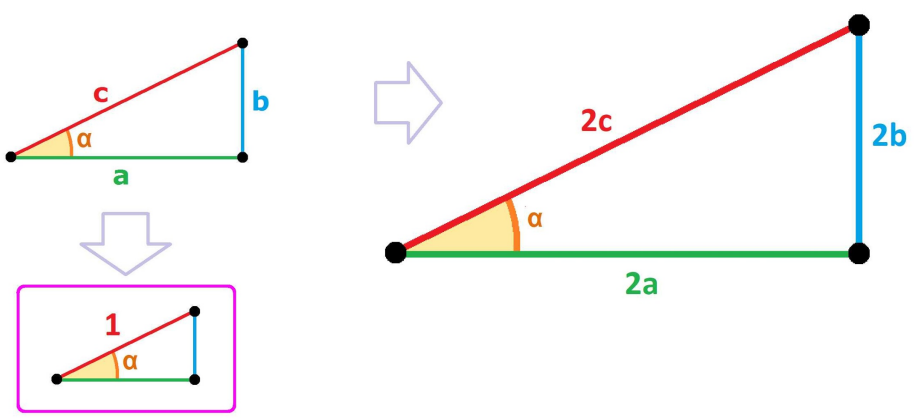
The importance of the tangent is seen in this formula:

$$\tan\alpha = \frac{b}{a} = \frac{\text{rise}}{\text{run}} = \textit{slope of the hypotenuse } c,$$

if side a follows the x-axis:



However, *similar triangles* have equal angles:



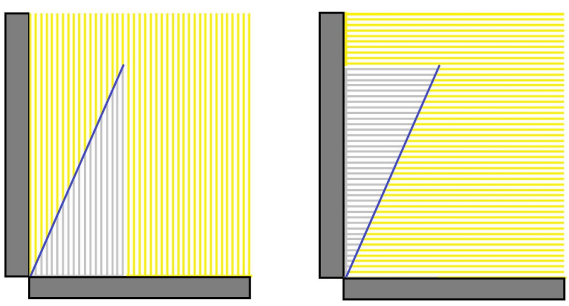
Then, let's pick the simplest, the one with a hypotenuse of length 1! Then our definitions don't need fractions anymore:

$$\begin{aligned}\cos\alpha &= a\\ \sin\alpha &= b\end{aligned}$$

But what does trigonometry have to do with periodicity and repetitiveness? Rotation.

Example 4.10.3: shadow

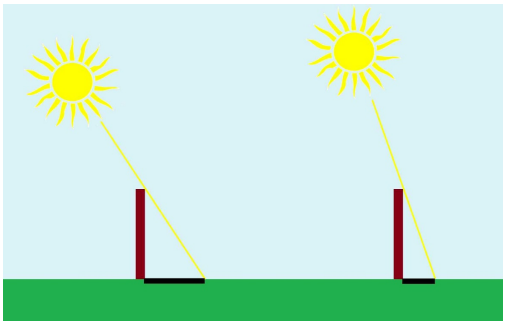
We have other interpretations of these quantities. Suppose we place a stick of length 1 at the angle α with the ground.



Then:

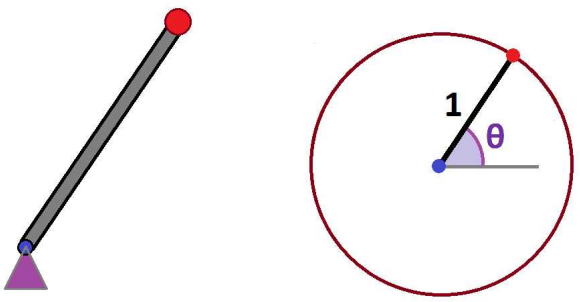
- We can think of $\cos\alpha$ as the length of its shadow – on the ground at noon (left).
- We can think of $\sin\alpha$ as the length of its shadow – on the wall at sunset (right).

Alternatively, the stick is vertical and still and it is the sun that is moving:

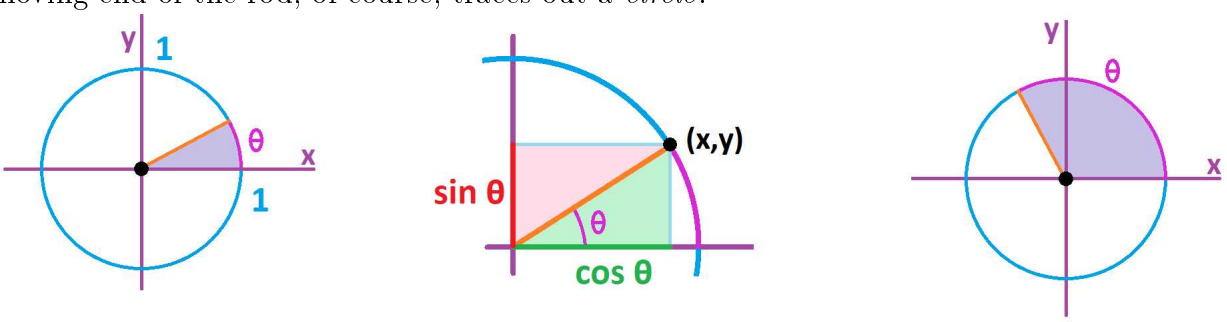


Example 4.10.4: rotating rod

Suppose a rod of length 1 is rotated around its end. If we can control the angle, θ, then what do we know about the *position* of its other end in space, i.e., its x and y coordinates?



The moving end of the rod, of course, traces out a *circle*:



We now look at the coordinates of the end point on this Cartesian plane. We use the above formulas:

$$x = \cos\theta \quad \text{and} \quad y = \sin\theta\,.$$

But do the formulas give us the *whole* circle? No. Because if we keep rotating beyond 90 degrees, there is – and there can be – no right triangle with this angle.

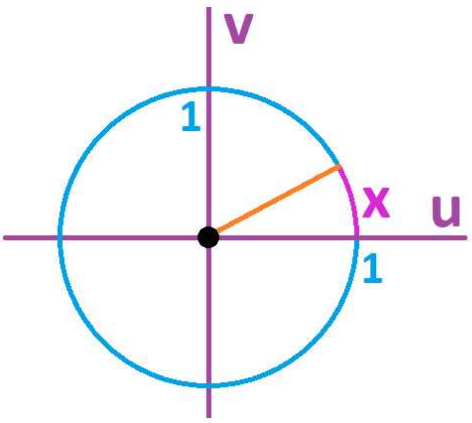
So, in a right triangle, an angle cannot go beyond 90 degrees. This isn't good enough anymore!

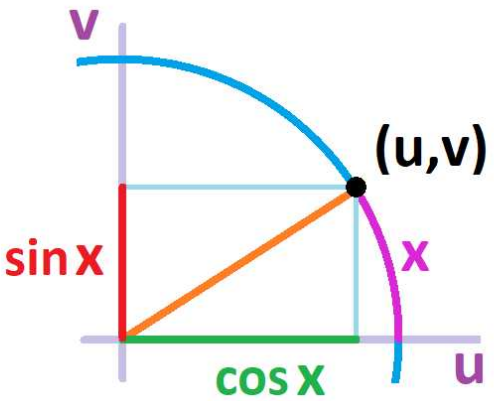
We need a new way to define the angle and the trigonometric functions of this angle. This is the idea:

▶ The angle isn't an angle of a triangle anymore but the angle of *rotation*.

First, we need to choose a name for the independent variable that represents the angle and runs throughout $(-\infty, +\infty)$. We call it, again, x.

We then are forced, second, to use alternative names for the axes of the coordinate plane where the construction is to happen; we call it the uv-plane:





Third, what units do we use for x? Practically, any unit of angle is acceptable: degrees, minutes, etc. In fact, the function they produce will differ only by a horizontal stretch (Chapter 3). However, just as e is the most natural choice of the base of the exponential function, there is the best choice for the unit of the angle x. The choice is based on the fact (to be demonstrated in Chapter 3IC-3) that the length of this unit circle is 2π and the length of its arc is proportional to this number and the angle it is based on. The length of the arc then can be used to measure the magnitude of the corresponding angle. In addition, the graph of the function $y = \sin x$ created this way – just like that of $y = e^x$ – crosses the y-axis at 45 degrees (Chapter 2DC-3)!

We adopt the **convention** that the size of the half-turn angle – and the length of the half-circle – is equal to π *radians*. Then, we have the following relation:

Definition 4.10.5: radians

180 degrees $= \pi$ radians.

We, therefore, have the following conversion formulas for these units:

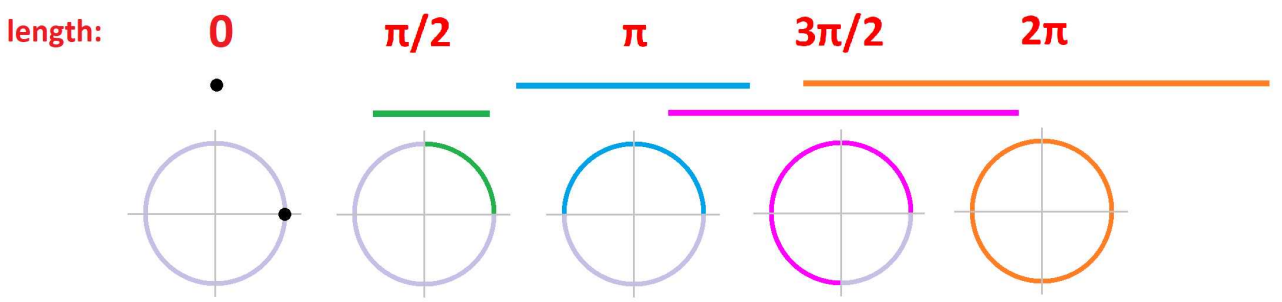
$$\#\text{ degrees} \quad = \quad \#\text{ of radians} \cdot \frac{180}{\pi}$$

The positive direction of the angle is *counterclockwise*:

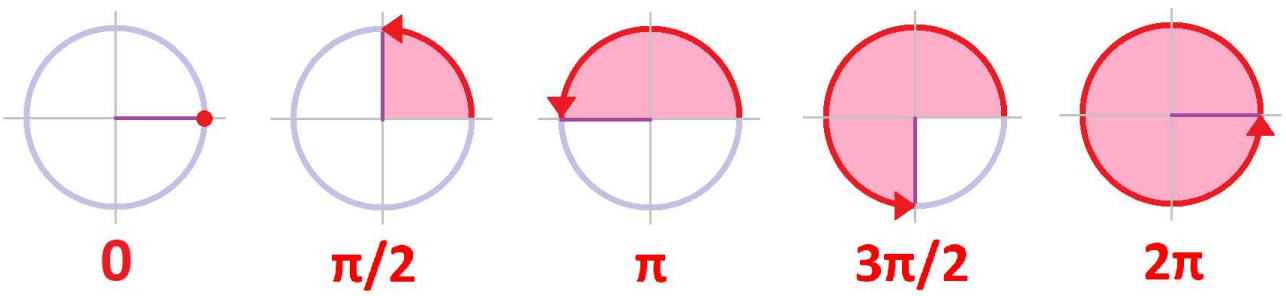
θ	degrees:	0	90	180	270	360	...
x	radians:	0	$\pi/2$	π	$3\pi/2$	2π	...

If the domain isn't the x-axis, can we still visualize it?

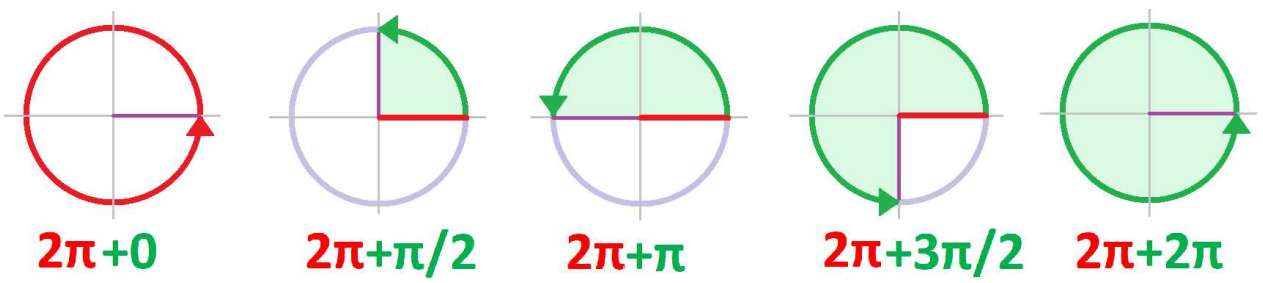
One way to see the x's is on the circle itself. To begin with, pieces of the interval $[0, 2\pi]$ are wrapped around this circle:



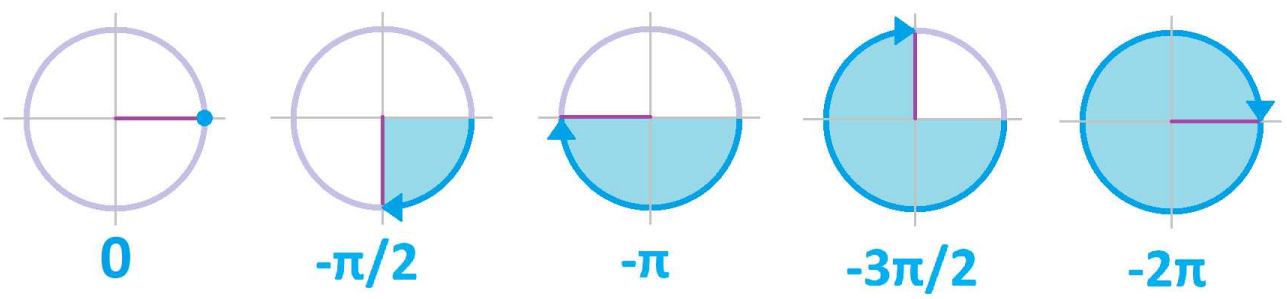
However, x is still the angle of rotation:



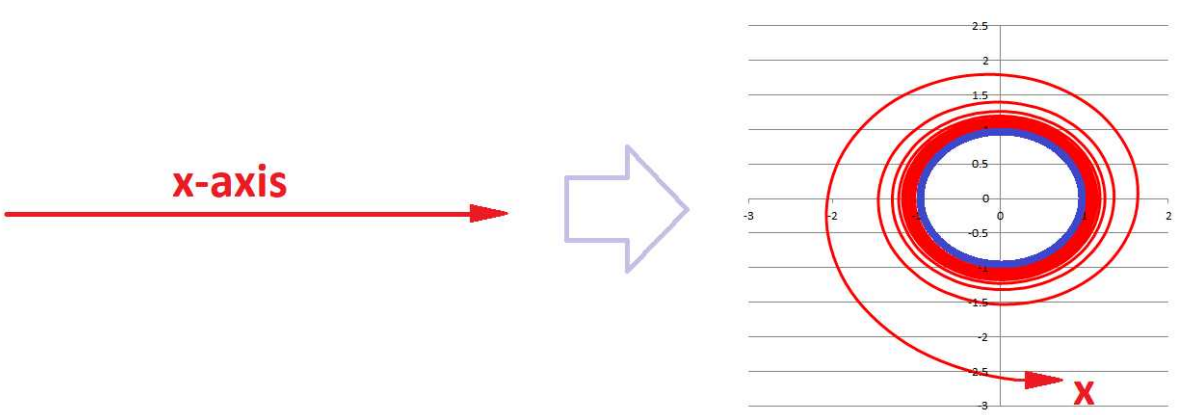
Then, of course, it can go past 360 degrees, like this:



That's clockwise; it can also go counterclockwise, with x negative:



The x-axis is then seen as a spiral wrapped on this unit circle, like this:



Finally, we are ready to construct these trigonometric functions. To accommodate all possible inputs, we will allow the outputs to be *negative* too; we use coordinates, as follows.

Definition 4.10.6: sine and cosine functions

Suppose a real number x is given. We construct a line segment of length 1 on the Cartesian plane (with the horizontal axis *not* marked x) starting at 0 with angle x radians from the horizontal, counterclockwise. Then:

- The *cosine* of x is the horizontal coordinate of the end of the segment.
- The *sine* of x is the vertical coordinate of the end of the segment.

They are denoted, respectively, by:

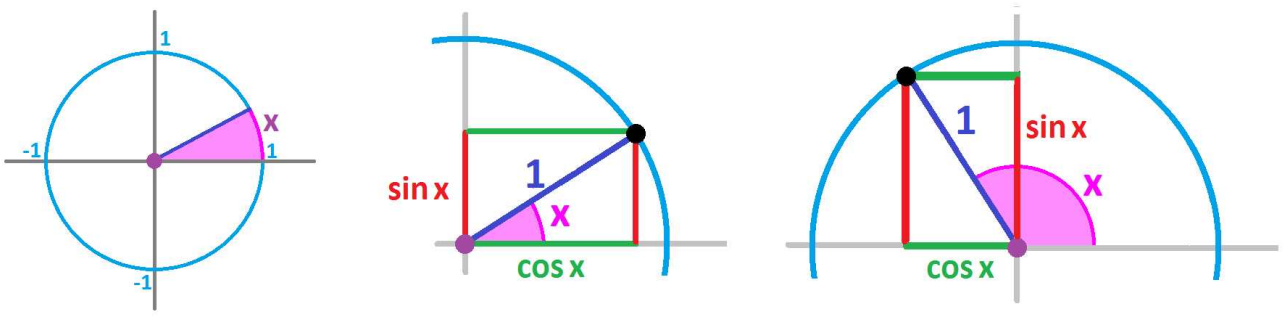
$$\cos x \quad \text{and} \quad \sin x$$

In other words, this is a two-step procedure:

$$x \to \boxed{\text{angle}} \to \boxed{\text{location on } uv\text{-plane}} \to \boxed{u\text{-coordinate}} \to y = \cos x$$

$$x \to \boxed{\text{angle}} \to \boxed{\text{location on } uv\text{-plane}} \to \boxed{v\text{-coordinate}} \to y = \sin x$$

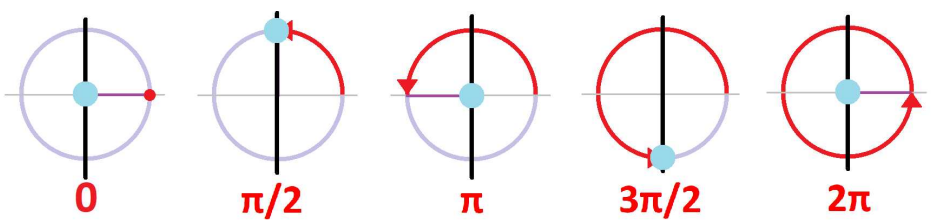
The outline of the construction is shown below:



Warning!

The "x" in $\cos x$ and $\sin x$ doesn't refer to the x-axis of the plane where the circle is plotted.

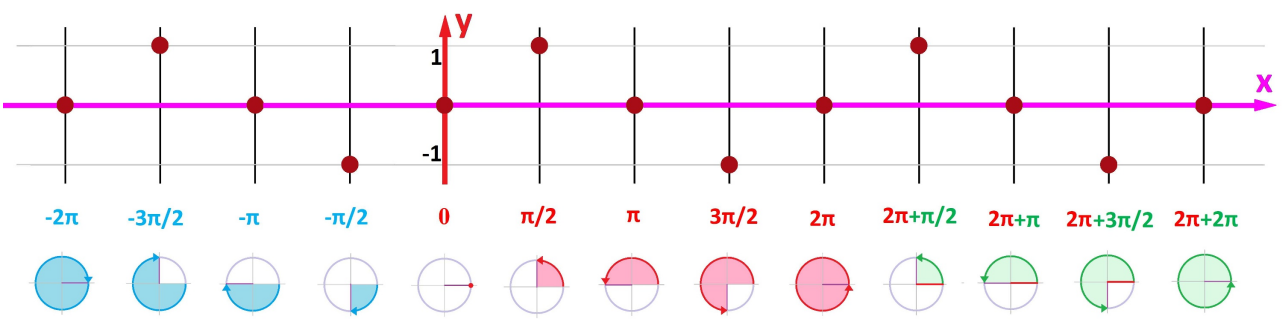
We can now find the values of $y = \sin x$ by examining the values of x, as angles, pictured above:



We get more and make a table:

input	x	...	-2π	$-3\pi/2$	$-\pi$	$-\pi/2$	0	$\pi/2$	π	$3\pi/2$	2π	...	horizontal
output	y	...	0	1	0	-1	0	1	0	-1	0	...	vertical

We plot the graph of $y = \sin x$ – point by point following the table above – with the domain presented as the x-axis on the xy-plane, as usual:



How do we fill the gaps in the graph? A method may be employed that is similar to the one we used for the exponential function:

► We divide intervals in half: if we know the function at two points, there is a formula to find the value in the middle.

Example 4.10.7: filling gaps in graphs

The trig formula we need is proven in the next chapter:

$$\sin\left(\frac{\alpha+\beta}{2}\right) = \pm\sqrt{\frac{1-\cos\alpha\cos\beta-\sin\alpha\sin\beta}{2}}\,.$$

As long as we know both the sine and the cosine of a couple of points, we can produce more and more.

For example, $\pi/4$ is half way between 0 and $\pi/2$. We just substitute these values from the table into the formula:

$$\begin{aligned}\sin\frac{\pi}{4} &= \sin\left(\frac{0+\pi/2}{2}\right)\\ &= \sqrt{\frac{1-\cos 0\cos\pi/2-\sin 0\sin\pi/2}{2}}\\ &= \sqrt{\frac{1-\cos 0\cos\pi/2-\sin 0\sin\pi/2}{2}}\\ &= \sqrt{\frac{1}{2}}.\end{aligned}$$

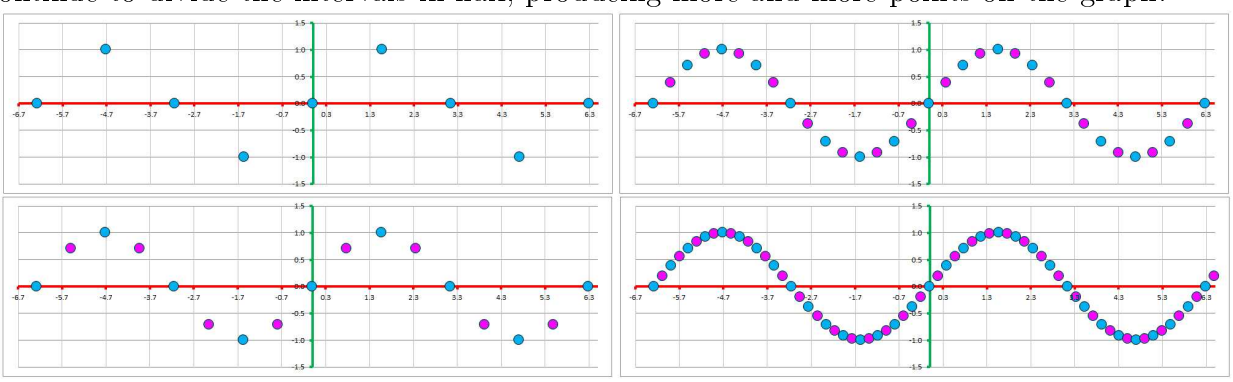
We compute the cosine with a similar formula.

Next, $\pi/8$ is half way between 0 and $\pi/4$. We just substitute the values we just found into the formula:

$$\sin\frac{\pi}{8} = \sin\left(\frac{0+\pi/4}{2}\right) = \ldots$$

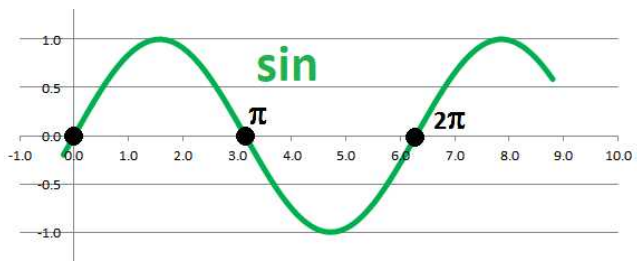
And so on.

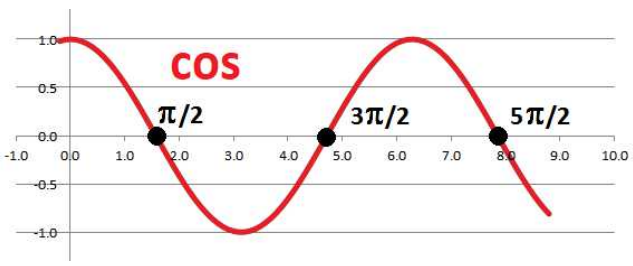
We continue to divide the intervals in half, producing more and more points on the graph:



The steps above, respectively, are: $\pi/2$, $\pi/4$, $\pi/8$ and $\pi/16$.

With more values found, these are the graphs of the sine and the cosine:





Exercise 4.10.8

What similarities between the two graphs do you see?

Exercise 4.10.9

Describe these functions as transformations.

Definition 4.10.10: tangent

The *tangent* function is defined by the following:

$$\tan x = \frac{\sin x}{\cos x}$$

Warning!

Following our conventions about functions, we can alternatively write: $\sin(x)$, $\cos(x)$, and $\tan(x)$, because, after all, sin, cos, and tan are the names of the functions and x is the independent variable.

To find the domain of a fraction, we set the denominator equal to 0 and solve. Find all x's that satisfy:

$$\cos x = 0\,.$$

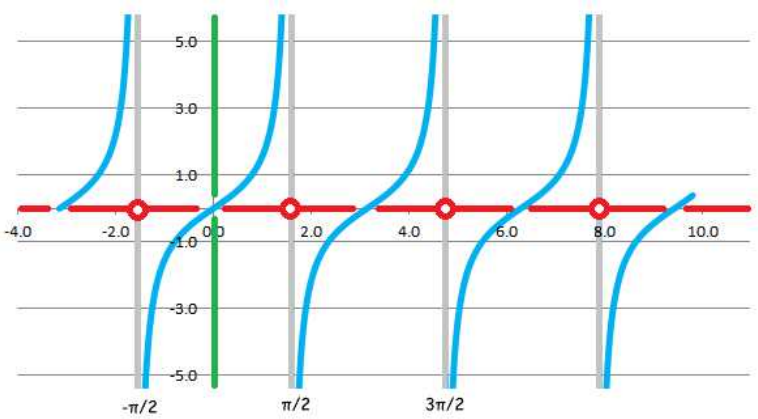
The solution set of this equation is what is excluded from the domain: the multiples of π starting from $\pi/2$.

Theorem 4.10.11: Domain of Tangent

The domain of the tangent, $y = \tan x$, is the following set:

$$\left\{x:\ x \neq \ldots,\ \frac{\pi}{2},\ \frac{3\pi}{2},\ \frac{5\pi}{2},\ \ldots\right\}.$$

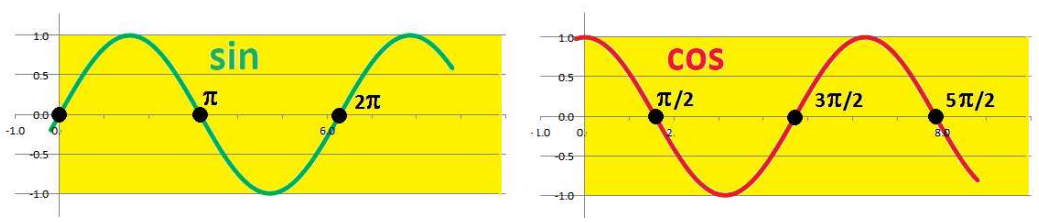
This is its graph:



Just as with the rational functions, the holes in the domain correspond to what appears to be vertical asymptotes.

Let's ask and answer some old questions about these new functions.

We take another look at the graphs of the two functions:



They fit inside a horizontal band! The following will re-appear many times in our study.

Theorem 4.10.12: Boundedness of Trigonometric Functions

The sine and the cosine are bounded functions with these bounds:

$$\begin{array}{rcccl} -1 & \leq & \sin x & \leq & 1 \\ -1 & \leq & \cos x & \leq & 1 \end{array}$$

satisfied for each x, while the tangent is unbounded.

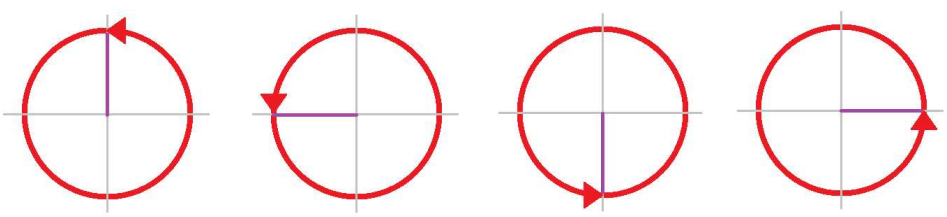
Exercise 4.10.13

Prove the last part.

Exercise 4.10.14

What are the images of these functions? What are the preimages?

After a full turn, we arrive to the same spot, no matter where we start:



But the values of the trigonometric functions are determined by this location only; that's 2π-periodicity. We will rely on the following:

Theorem 4.10.15: Periodicity of Trigonometric Functions

The functions sin *and* cos *are periodic with period 2π, while the function* tan *is periodic with period π; i.e.,*

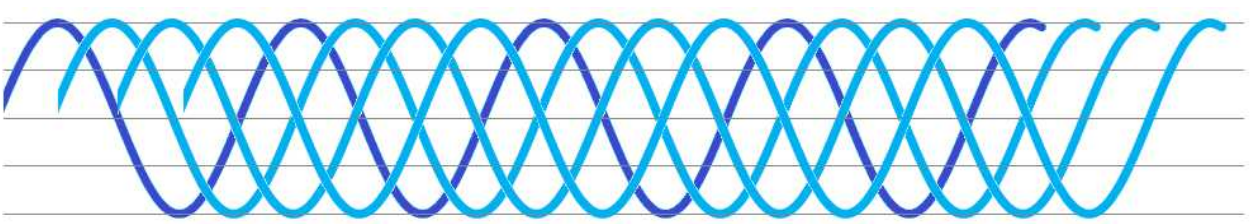
$$\begin{array}{rcl} \sin(x+2\pi) & = & \sin x \\ \cos(x+2\pi) & = & \cos x \\ \tan(x+\pi) & = & \tan x \end{array}$$

satisfied for each x.

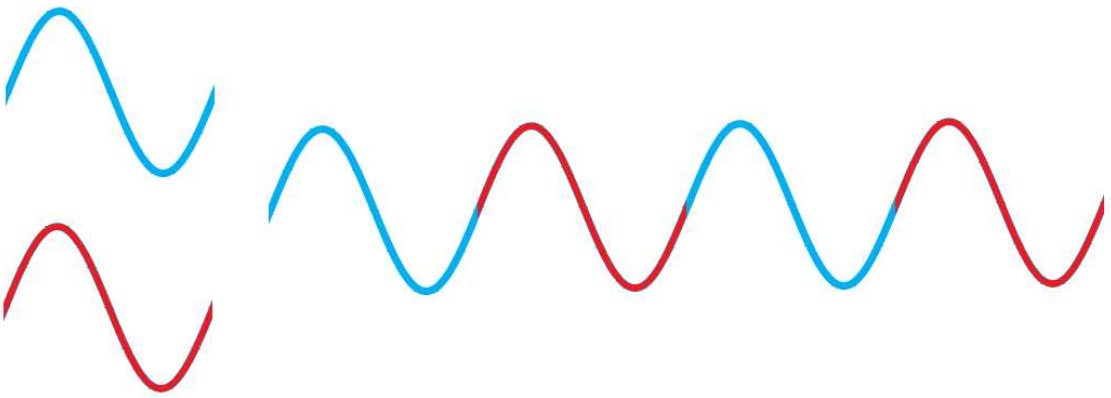
Exercise 4.10.16

Prove the last part.

One can see the periodicity in the graphs. There are two ways. First is the shift:

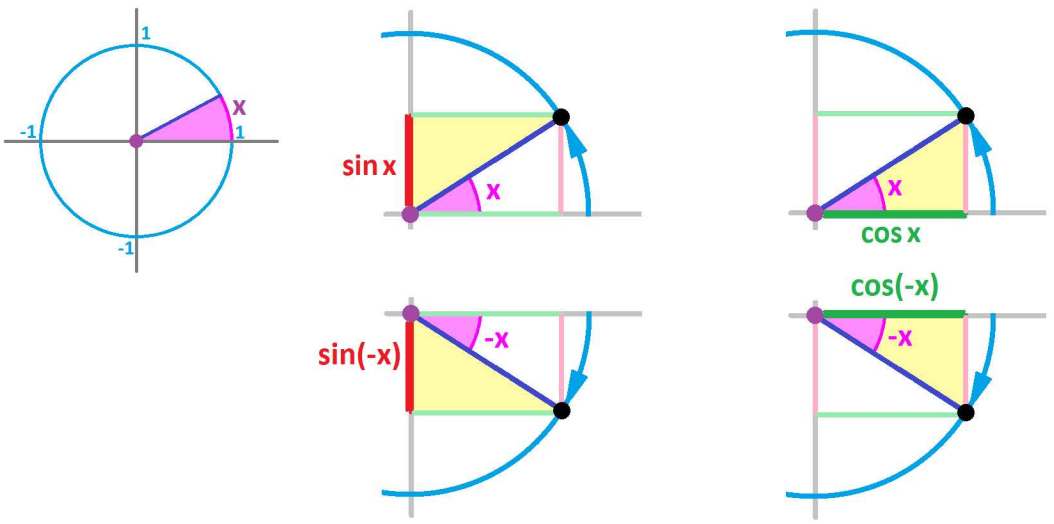


The second is copy-and-paste:



Now, what about the *symmetries*?

Le's rotate our rod the same amount, x, clockwise and counter-clockwise and compare:



The effect is different for the two functions:

1. It makes the vertical progress, $\sin x$, negative.
2. It doesn't change the horizontal progress, $\cos x$.

The following is a convenient observation.

Theorem 4.10.17: Odd-Even Trig Functions

The sine and the tangent are odd functions, while the cosine is even.

In other words, we have:

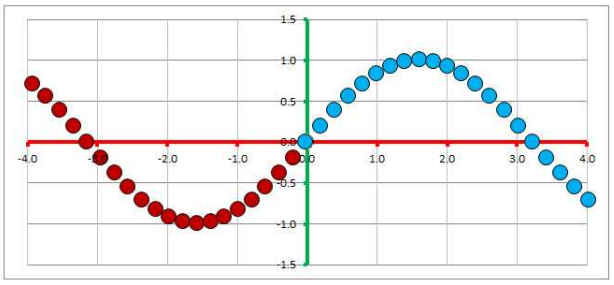
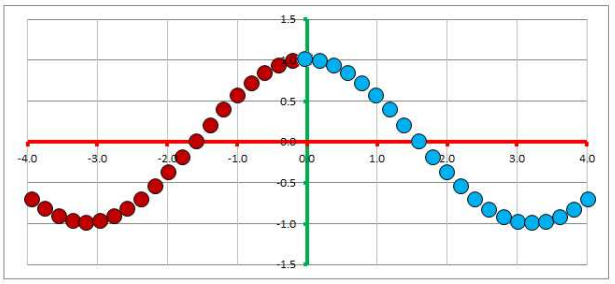
$$\begin{aligned} \sin(-x) &= -\sin x \\ \cos(-x) &= \cos x \\ \tan(-x) &= -\tan x \end{aligned}$$

satisfied for each x.

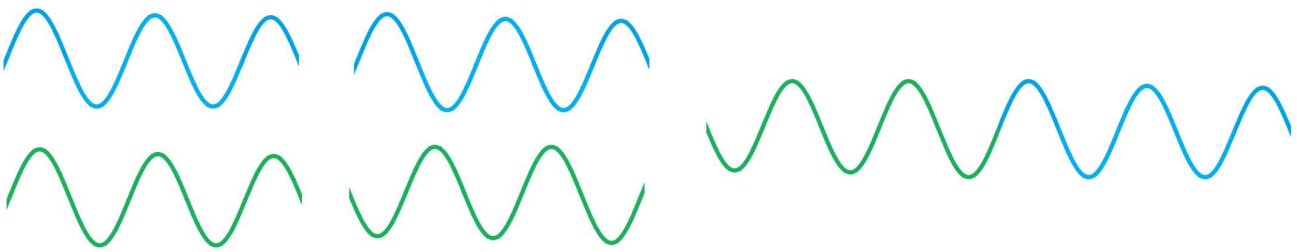
Exercise 4.10.18

Prove the part about tangent.

The graphs do exhibit these symmetries:

We can see this fact about the sine via copy-and-paste:



In fact, when zoomed in on the y-intercept, the graph of $y = \sin x$ looks like $y = x$ and $y = \cos x$ like $y = 1$!

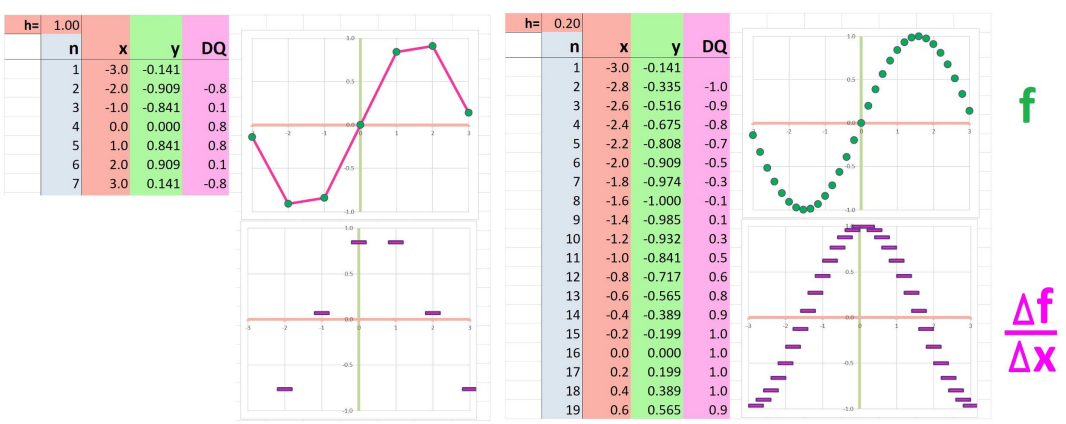
Exercise 4.10.19

What parabola does the graph of the cosine look like around the y-intercept?

Exercise 4.10.20

What does the the graph of the tangent look like if we zoom in on it around the origin?

This is the difference quotient $\dfrac{\Delta f}{\Delta x}$, i.e., the sampled slopes, of $\sin x$ (bottom row):



Exercise 4.10.21

What pattern do the slopes exhibit?

Exercise 4.10.22

Plot the difference quotient for cosine and answer the same question.

None of these functions is one-to-one.

None, therefore, can have an inverse.

However, what we did with $y = x^2$ in order to get $x = \sqrt{y}$ can be repeated here:

- ▶ Restrict the domains of the function to make it one-to-one.

The codomain is also restricted so that the function is also onto.

Theorem 4.10.23: Restricted Sine

- *With the domain restricted to* $[-\pi/2, \pi/2]$*, the sine,* $y = \sin x$*, is a one-to-one function.*
- *With the codomain restricted to* $[-1, 1]$*, the sine,* $y = \sin x$*, is an onto function.*

Why *this* interval for the domain? First, it is, in a sense, the largest possible: we can't extend the interval – to the left or to the right – without making the function not one-to-one. But, second, when restricted this way, the sine is still an odd function. And now so is its inverse.

Definition 4.10.24: arcsine

The *arcsine* is defined to be the inverse of the sine function restricted to $[-\pi/2, \pi/2]$, denoted by either:

$$\arcsin y = \sin^{-1} y$$

Warning!

Just as there is no "formula" for the logarithm or for the square root, there is none for arcsin.

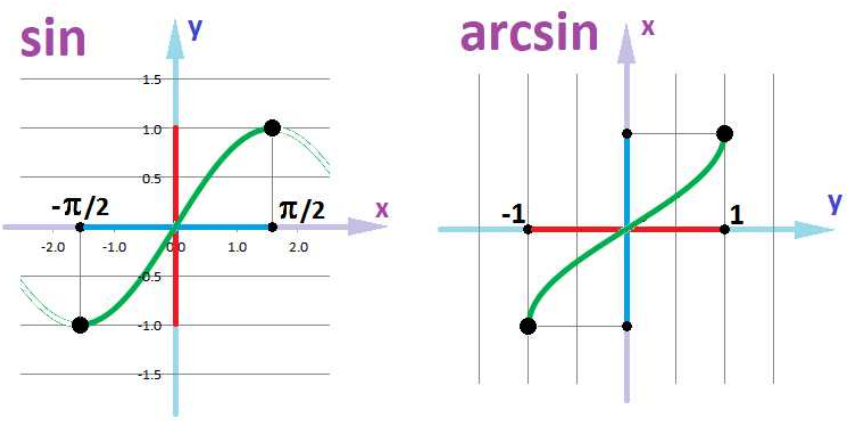
Exercise 4.10.25

Show that they are both *increasing*.

Thus, we have a pair of inverse functions:

$y = \sin x$	$x = \sin^{-1} y$
domain: $[-\pi/2, \pi/2]$	domain: $[-1, 1]$
range: $[-1, 1]$	range: $[-\pi/2, \pi/2]$

The graphs are, of course, the same with just x and y interchanged:



Exercise 4.10.26

Prove that arcsin is odd.

Similarly, we choose the to restrict the domain of the new cosine to the interval $[0, \pi]$. Why *this* interval? First, we can't extend beyond this interval because that would make the function not one-to-one. But second, we take the half of the graph that is repeated on the other side of the y-axis because the cosine is even.

Theorem 4.10.27: Restricted Cosine

- *With the domain restricted to* $[0, \pi]$*, the cosine,* $y = \cos x$*, is a one-to-one function.*
- *With the codomain restricted to* $[-1, 1]$*, the cosine,* $y = \cos x$*, is an onto function.*

Definition 4.10.28: arccosine

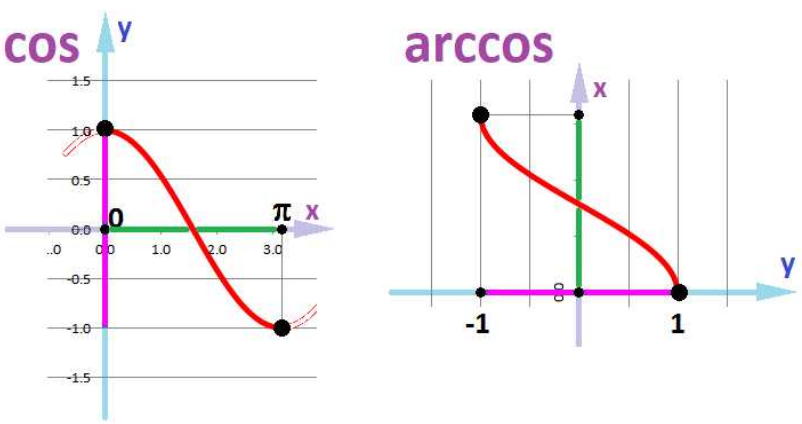
The *arccosine* is defined to be the inverse of the cosine function on $[0, \pi]$, denoted by either:

$$\arccos y = \cos^{-1} y$$

Thus, we have a pair of inverse functions:

$y = \cos x$	$x = \cos^{-1} y$
domain: $[0, \pi]$	domain: $[-1, 1]$
range: $[-1, 1]$	range: $[0, \pi]$

The graphs are, of course, the same with just x and y interchanged:

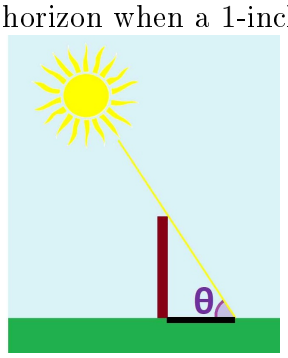


Exercise 4.10.29

Show that both sine and cosine are one-to-one on any interval of length π.

Example 4.10.30: angle of sun

Armed only with this knowledge and a calculator, we can try to solve some practical problems. For example, how high is the sun above the horizon when a 1-inch stick casts a shadow .5 inch long?



We measure the height of the sun in terms of the angle x the sunlight hits the ground. Then, we have the following:

$$\cos x = 1/2 \implies x = \arccos(1/2) \approx 1.05\,.$$

This is, in fact, 60 degrees.

The tangent function is π-periodic and, therefore, not one-to-one. To build its inverse, we, again, restrict the function's domain. We simply choose the branch of tan over $-\pi/2 < x < \pi/2$.

Theorem 4.10.31: Restricted Tangent

With the domain restricted to $(-\pi/2, \pi/2)$*, the tangent,* $y = \tan x$*, is a one-to-one function.*

It's still odd and so is its inverse.

Definition 4.10.32: arctangent

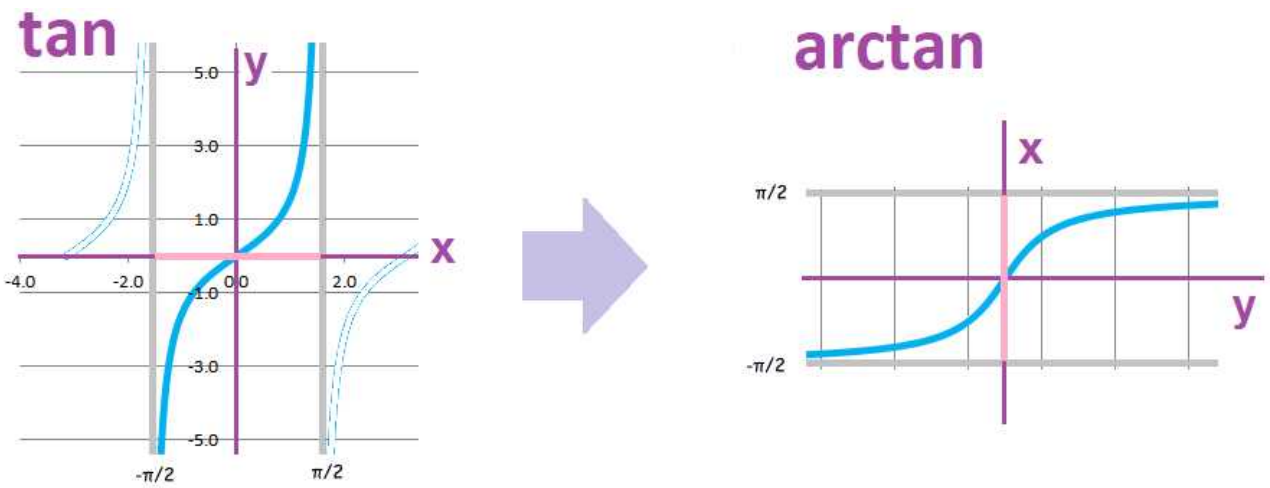
The *arctangent* is defined to be the inverse of the tangent function restricted to the domain $(-\pi/2, \pi/2)$, denoted by either:

$$\arctan y = \tan^{-1} y$$

Then, we again have a pair of inverse functions:

$y = \tan x$	$x = \tan^{-1} y$
domain: $(-\pi/2, \pi/2)$	domain: $(-\infty, +\infty)$
range: $(-\infty, +\infty)$	range: $(-\pi/2, \pi/2)$

The graphs are, of course, the same with just x and y interchanged:



We use these functions to solve *trigonometric equations*.

Example 4.10.33: solving trig equations

Solve the following for x:

$$\sin(x - 5) = .3\,.$$

To "kill" the sine and get to x, apply its *inverse*, the arcsine, to both sides of the equation:

$$\arcsin\big(\sin(x - 5)\big) = \arcsin(.3)\,.$$

We cancel as before and finish:

$$\begin{aligned} x - 5 &= \arcsin .3 \qquad \Longrightarrow \\ x &= \arcsin .3 + 5\,. \end{aligned}$$

Exercise 4.10.34

Have we found all solutions?

Below, we summarize how, hypothetically, these classes of functions could have appeared:

History of functions

Phenomena	Requirements for new functions	Inverses
rolling ball	Needs to vary.	
	$\longrightarrow$ linear functions	$\longrightarrow$ linear functions
thrown ball	Needs to change at a variable rate.	
	$\longrightarrow$ polynomials	$\longrightarrow$ "algebraic functions"
gravity in space	Needs to gradually diminish.	
	$\longrightarrow$ rational functions	$\longrightarrow$ "algebraic functions"
population growth/decline, cooling/heating	Needs to grow and decline faster.	
	$\longrightarrow$ exponential functions	$\longrightarrow$ logarithms
waves, planetary motion	Needs to repeat itself.	
	$\longrightarrow$ trig functions	$\longrightarrow$ inverse trig functions
new, unknown phenomena	Will need to conform.	
	$\longrightarrow$ abstract functions	

Chapter 5: Algebra and geometry

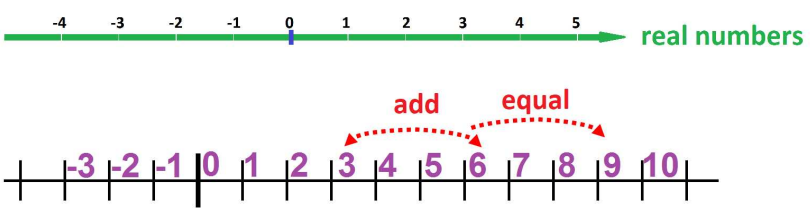
Contents

5.1 The arithmetic operations on functions 394
5.2 The algebra of compositions 402
5.3 Solving equations 409
5.4 The algebra of logarithms 418
5.5 The Cartesian system for the Euclidean plane 429
5.6 The Euclidean plane: distances 437
5.7 Trigonometry and the wave function 451
5.8 The Euclidean plane: angles 465
5.9 From geometry to calculus 475
5.10 Solving inequalities 480

5.1. The arithmetic operations on functions

We would like to treat *all numerical functions as a single group.*

We find inspiration in how we have handled the *real numbers.* We put them together in the real number line, which provides us with a bird's-eye view:



We also recognize that these entities are interacting with each other, producing offspring via arithmetic:

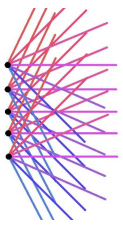
$$3+6=9,\ 5\cdot 7=35,\ \text{etc.}$$

Understanding the meaning of these computations requires understanding that the beginning and the end of such a computation are just two different representations of the *same* number:

$$1+1\ =\ 2\cdot 1\ =\ 2\,.$$

There is a single location for each of these expressions on the real number line! Similarly, $x+x$ and $2x$ correspond to the same (albeit unspecified) location on the number line. We simply say that they are *equal.*

It is much more challenging to find such a bird's-eye view for *functions*! For example, this is what an attempt to visualize all *linear* functions would look like:



Just as with numbers, it is *interactions* between functions that make them manageable as a whole.

For each of the four arithmetic operations on *numbers* – addition, subtraction, multiplication, and division – there is an operation on (numerical) *functions*.

But first let's make sure that we have a clear understanding of what it means for two functions to be the same. For example, these two functions are represented by two different formulas:

$$x + x \quad \text{and} \quad 2x\,.$$

Are they the same function? Of course! These two functions are represented by two similar formulas:

$$x - 1 \quad \text{and} \quad 1 - x\,.$$

Are they the same function? Of course not! How do we know?

The answer (as is the question itself) is dependent on our definition of function:

A function is a list of inputs and outputs.

To answer the question, we can simply test the formulas by plugging input values and watching the outputs:

$$\begin{array}{lll} x + x\Big|_{x=0} = 0, & 2x\Big|_{x=0} = 0 & \text{same!} \\ x + x\Big|_{x=1} = 2, & x + x\Big|_{x=1} = 2 & \text{same!} \\ \ldots & \ldots & ??? \end{array}$$

It seems the same. Now the second pair:

$$\begin{array}{lll} x - 1\Big|_{x=0} = -1, & 1 - x\Big|_{x=0} = 1 & \text{different!} \\ x - 1\Big|_{x=1} = 0, & 1 - x\Big|_{x=1} = 0 & \text{same!} \end{array}$$

We stop here because a single mismatch means that they are different!

But what about these functions:

$$\frac{2x^2 + 2x}{2} \quad \text{and} \quad x^2 + x\,,$$

or those:

$$\frac{2x^2 + 2x}{x} \quad \text{and} \quad 2x + 2\,?$$

We again plug in the values:

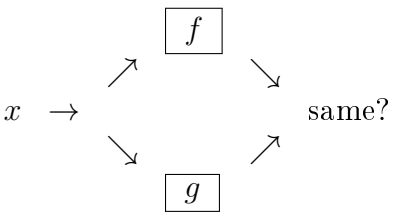
$$\begin{array}{lll} \dfrac{2x^2 + 2x}{2}\Big|_{x=0} = 0, & x^2 + x\Big|_{x=0} = 0 & \text{same!} \\ \dfrac{2x^2 + 2x}{2}\Big|_{x=1} = 2, & x^2 + x\Big|_{x=1} = 2 & \text{same!} \\ \ldots & \ldots & ??? \end{array}$$

The results are the same for every x! What about the latter? It breaks down:

$$\left.\frac{2x^2+2x}{x}\right|_{x=0} \text{ undefined,} \quad \left.2x+2\right|_{x=0} = 2 \quad \text{different!}$$

Plugging in $x = 0$ will produce division by 0 for the first function in the pair but not for the second. It is clear then that two functions can't be the same unless their domains are equal too (as sets).

The following is the test two function functions f and g are subjected to:



So, f and g are called *equal*, or we say it's *the same function*, if they have the same domain and

$$f(x) = g(x) \quad \texttt{FOR EACH } x$$

in the domain.

These are our answers to the above questions. Are these two functions the same:

$$f(x) = \frac{2x^2+2x}{2} \quad \text{and} \quad g(x) = x^2 + x \ ?$$

Yes, because

$$\frac{2x^2+2x}{2} = x^2 + x \quad \text{for every } x.$$

It is crucial that the *implied domains* – for every x – of the two functions are the same.

We use the same simple notation for functions as for numbers:

Equal functions

$$f = g$$

Are these two functions the same:

$$f(x) = \frac{2x^2+2x}{x} \quad \text{and} \quad g(x) = 2x + 2 \ ?$$

No, because the implied domain of the former doesn't include 0 while that of the latter does. The difference is in a single value!

We also use this simple notation:

Not equal functions

$$f \neq g$$

As you can see, once we discover that the domains don't match, we are done. However, choosing another domain will fix the problem:

$$f(x) = \frac{2x^2+2x}{x} \quad \text{and} \quad g(x) = 2x + 2 \quad \text{are the same function on the domain } \{x : \ x \neq 0\}.$$

As another relevant example, these are two different functions:

- x^2 with domain $(-\infty, \infty)$;
- x^2 with domain $[0, \infty)$.

Exercise 5.1.1

Consider:

$$\frac{x}{x^2} \quad \text{vs.} \quad \frac{1}{x}\,.$$

Exercise 5.1.2

Suggest your own examples of functions that differ by a single value.

The statement in the definition, such as

$$\frac{2x^2+2x}{2} = x^2+x \quad \text{for every real } x,$$

is called an *identity*. The last part is often assumed and omitted from computations. The following statement is also an identity:

$$\frac{2x^2+2x}{x} = 2x+2 \quad \text{for every real } x \neq 0\,.$$

However, the last part is a caveat that *cannot* be omitted! In other words, an identity is just a statement about two functions being "identically" equal, i.e., indistinguishable, within the specified domain.

This idea of transitioning from a function to its *twin* is the basis of all algebraic manipulations; they are informally called "simplifications" or "cancellations".

Now, the outputs of numerical functions are *numbers*. Therefore, any arithmetic operation on numbers – addition, subtraction, multiplication, and division – can now be applied to functions, one input at a time. Once again, functions interact and produce offspring, new functions.

The definitions of these new functions are simple.

Definition 5.1.3: sum of functions

Given two functions f and g, the *sum*, $f+g$, of f and g is the function defined by the following:

$$(f+g)(x) = f(x)+g(x) \quad \texttt{FOR EACH } x$$

in the intersection of the domains of f and g.

Note how the two plus signs in the formula are different: The first one is a part of the *name* of the new function while the second is the actual sign of summation of two real numbers. This is the "deconstruction" of the notation:

Sum of functions

$$\underset{\substack{\uparrow\\ \text{name of the new function}}}{\big(\ f+g\ \big)\ (x)} \quad = \quad \underset{\substack{\text{names of the first and second functions}\\ \downarrow \qquad \downarrow\\ f(x)+g\ (x)\\ \uparrow\ \uparrow\ \uparrow\\ \text{operation on numbers}}}{}$$

Furthermore, we now have an operation on functions: $f+g$ is a new function.

Example 5.1.4: algebra of functions

The sum of

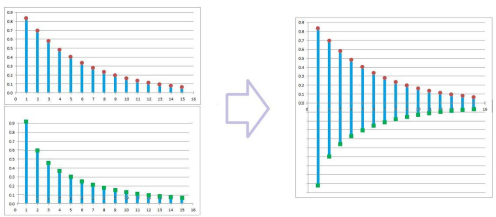
$$g(x) = x^2 \quad \text{and} \quad f(x) = x+2$$

is

$$(g+f)(x) = g(x) + f(x) = \left(x^2\right) + \left(x+2\right).$$

Whether this is to be simplified or not, a new function has been built.

This is an illustration of the meaning of the sum of two functions:



One can see how the values are added, location by location.

We represent a function f diagrammatically as a *black box* that processes the input and produces the output:

input		function		output
x	$\to$	$\boxed{f}$	$\to$	y

Now, suppose we have another function g:

input		function		output
t	$\to$	$\boxed{g}$	$\to$	u

How do we represent their sum $f+g$? To represent it as a single function, we need to "wire" their diagrams together side by side:

$$\begin{array}{ccccc} x & \to & \boxed{f} & \to & y \\ \| & & & & \updownarrow \\ t & \to & \boxed{g} & \to & u \end{array}$$

But it's only possible when the input of f coincides with the input of g. We may have to *rename the variable* of g. We replace t with x. Then we have a new diagram for a new function:

$$f+g: \quad x \to \quad x \quad \begin{array}{ccccc} \nearrow & x & \to & \boxed{f} & \to & y & \searrow \\ \searrow & x & \to & \boxed{g} & \to & u & \nearrow \end{array} \quad \boxed{\text{add}} \to z \to z$$

We see how the input variable x is copied into the two functions, processed by them *in parallel*, and finally the two outputs are added together to produce a single output. The result can be seen as just a new black box:

$$x \to \boxed{f+g} \to y$$

Warning!

When units are involved, we must make sure that the outputs match so that we can add them.

Example 5.1.5: algebra of functions

We have combined two functions into one but we often need to go the other way and break a complex function into simpler parts that can then be studied separately. Represent

$$z = h(x) = x^2 + \sqrt[3]{x}$$

as the sum of two functions. Here is the answer:

$$x \mapsto y = x^2 \quad \text{and} \quad x \mapsto y = \sqrt[3]{x}\,.$$

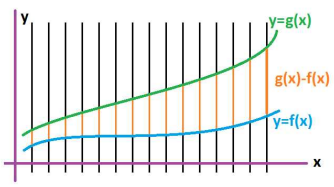
Subtraction also gives us an operation on functions.

Definition 5.1.6: difference of functions

Given two functions f and g, the *difference*, $g - f$, of f and g is the function defined by the following:

$$(g - f)(x) = g(x) - f(x) \quad \texttt{FOR EACH } x$$

in the intersection of the domains of f and g.



Before we get to multiplication of functions, there is a simpler but very important version of this operation.

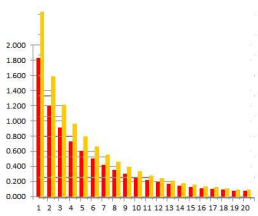
Definition 5.1.7: constant multiple of function

Given a function f, the *constant multiple* cf of f, for some real number c, is the function defined by the following:

$$(cf)(x) = cf(x) \quad \texttt{FOR EACH } x$$

in the domain of f.

In the following illustration of the meaning of a constant multiple of a function, one can see how its values are multiplied by $c = 1.3$ one location at a time:



There may be more than two functions involved in these operations or they can be combined.

Example 5.1.8: algebra of functions

Sum combined with differences:

$$h(x) = 2x^3 - \frac{5}{x} + 3x - 4\,.$$

The function is also seen as the sum of constant multiples, called a "linear combination":

$$h(x) = 2 \cdot \left(x^3\right) + (-5) \cdot \frac{1}{x} + 3 \cdot x + (-4) \cdot 1\,.$$

Example 5.1.9: algebra of functions given by tables

When two functions are represented by their lists of values, their sum (difference, etc.) can be easily computed. We simply go row by row adding the values.

Suppose we need to add these two functions, f and g, and create a new one, h, represented by a similar

list:

x	$y=f(x)$
0	1
1	2
2	3
3	0
4	1

$+$

x	$y=g(x)$
0	5
1	-1
2	2
3	3
4	0

$=$?

We simply add the output of the two functions for the same input. First row:

$$f:0\mapsto 1,\quad g:0\mapsto 5\quad\Longrightarrow\quad h:0\mapsto 1+5=6\,.$$

Second row:

$$f:1\mapsto 2,\quad g:1\mapsto -1\quad\Longrightarrow\quad h:1\mapsto 2+(-1)=1\,.$$

And so on. This is the whole solution:

x	$y=f(x)$
0	1
1	2
2	3
3	0
4	1

$+$

x	$y=g(x)$
0	5
1	-1
2	2
3	3
4	0

$=$

x	$y=f(x)+g(x)$
0	$1+5=6$
1	$2+(-1)=1$
2	$3+2=5$
3	$0+3=3$
4	$1+0=1$

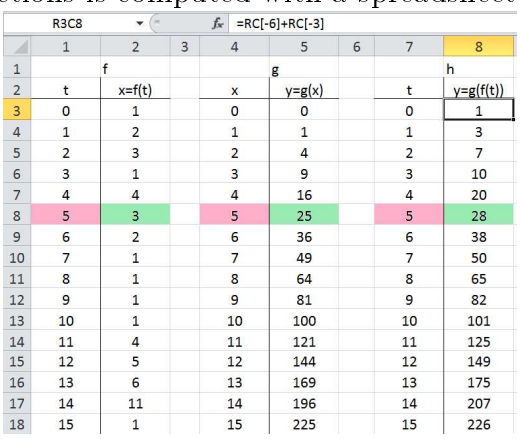
$=$

x	$y=h(x)$
0	6
1	1
2	5
3	3
4	1

.

Example 5.1.10: algebra of functions with spreadsheet

This is how the sum of two functions is computed with a spreadsheet:



	1	2	3	4	5	6	7	8
1		f			g			h
2	t	x=f(t)		x	y=g(x)		t	y=g(f(t))
3	0	1		0	0		0	1
4	1	2		1	1		1	3
5	2	3		2	4		2	7
6	3	1		3	9		3	10
7	4	4		4	16		4	20
8	5	3		5	25		5	28
9	6	2		6	36		6	38
10	7	1		7	49		7	50
11	8	1		8	64		8	65
12	9	1		9	81		9	82
13	10	1		10	100		10	101
14	11	4		11	121		11	125
15	12	5		12	144		12	149
16	13	6		13	169		13	175
17	14	11		14	196		14	207
18	15	1		15	225		15	226

The formula is very simple:

```
=RC[-6]+RC[-3]
```

There are two more operations, multiplication and division.

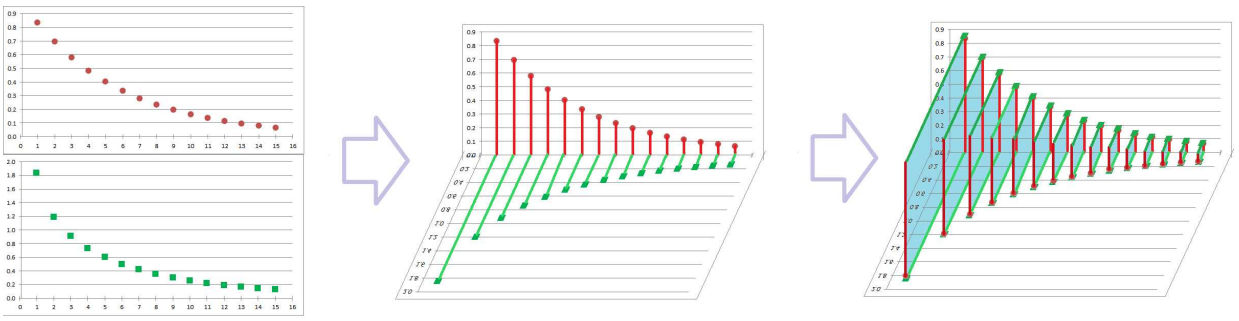
Definition 5.1.11: product of functions

Given two functions f and g, the *product*, $f\cdot g$, of f and g is the function defined by the following:

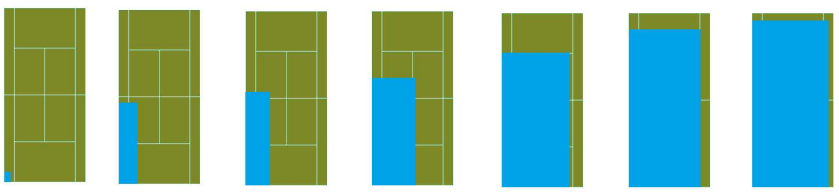
$$(f\cdot g)(x)=f(x)\cdot g(x)\quad\texttt{FOR EACH}\ x$$

in intersection of the domains of f and g.

For each value of x, we can use the pair $f(x)$ and $g(x)$ as the sides of a rectangle. Then the product $f(x)\cdot g(x)$ is seen as the area of this rectangle:



If we think of x as time, we can see this construction as a short clip of spreading a tarp:



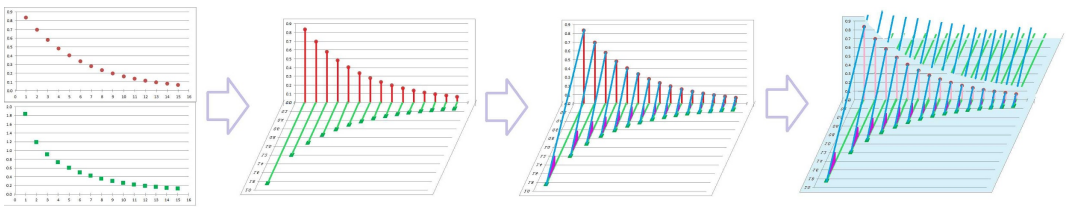
Definition 5.1.12: quotient of functions

Given two functions f and g, the *quotient*, f/g, of f and g is the function defined by the following:

$$\big(f/g\big)(x) = f(x)/g(x) \quad \texttt{FOR EACH } x$$

in the intersection of the domains of f and g with $g(x) \neq 0$.

For each value of x, we can use the pair $f(x)$ and $g(x)$ as the sides of a right triangle. They are horizontal and vertical, respectively, the run and the rise. Then the quotient $f(x)/g(x)$ is the *slope* of this line:

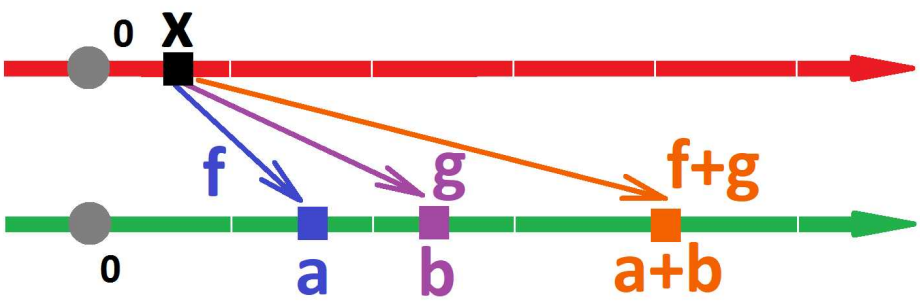


Exercise 5.1.13

Explain the difference between these two functions:

$$\sqrt{\frac{x-1}{x+1}} \quad \text{and} \quad \frac{\sqrt{x-1}}{\sqrt{x+1}}\,.$$

In summary, the outputs are on the y-axis and its algebra allows us to find the outputs for the sum and other algebraic operations on f, g:



All four algebraic operations produce new functions in the same manner:

$$x \to \quad x \quad \begin{array}{l} \nearrow x \to \boxed{f} \to y \searrow \\ \searrow x \to \boxed{g} \to u \nearrow \end{array} \quad \boxed{+\ -\ \cdot\ \div} \to z$$

Exercise 5.1.14

Have we found the domains of these new functions?

Warning!

The algebra of functions comes from the algebra of *outputs*; the inputs don't even have to be numbers.

Composition, however, is the most important operation on functions. There is no matching operation for numbers.

5.2. The algebra of compositions

Functions can be visualized as *flowcharts* and so can their compositions:



If we name the variables and use the algebraic notation, we produce a more compact version of this flowchart:

$$x \to \boxed{x+3} \to y \to \boxed{y \cdot 2} \to z \to \boxed{z^2} \to u$$

Note how the names of the variables match so that we can proceed to the next step. A purely algebraic representation of the diagram is below:

$$x+3=y, \quad y \cdot 2 = z, \quad z^2 = u\,.$$

It is also possible, but not required, to *name the functions*, say f, g, h. Then we have:

$$y = f(x) = x+3, \quad z = g(y) = y \cdot 2, \quad u = h(z) = z^2\,.$$

As we see, with the variables properly named,

composition is substitution.

In the above composition, we can carry out these two substitutions:

- We substitute $z = g(y) = y \cdot 2$ into $u = h(z) = z^2$, which results in the following:

$$u = h(z) = h(g(y)), \quad u = z^2 = (y \cdot 2)^2\,.$$

- We substitute $y = f(x) = x+3$ into $z = g(y) = y \cdot 2$, which results in the following:

$$z = g(y) = g(f(x)), \quad z = y \cdot 2 = (x+3) \cdot 2\,.$$

In general, we represent a function f diagrammatically as a *black box* that processes the input and produces the output:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ x & \to & \boxed{f} & \to & y \end{array}$$

Now, suppose we have another function g:

$$\begin{array}{ccccc} \text{input} & & \text{function} & & \text{output} \\ x & \to & \boxed{g} & \to & y \end{array}$$

How do we represent their composition $g \circ f$? To represent it as a single function, we need to "wire" their diagrams together *consecutively* (instead of in parallel, as in the last section):

$$x \to \boxed{f} \to y \to ??? \to x \to \boxed{g} \to y$$

But it's only possible when the output of f matches with the input of g. We can *rename the variable* of g. For example, we can make this switch:

$$\frac{x^2-1}{x+2} \to \frac{y^2-1}{y+2}.$$

Warning!

If the names of the variables don't match, it might be for a good reason.

This is what we have after renaming:

$$x \to \boxed{f} \to y \to \boxed{g} \to z$$

Then we have a new diagram for a new function:

$$g \circ f: \quad x \to \boxed{x \to \boxed{f} \to y \to \boxed{g} \to z} \to z$$

It's just another black box:

$$x \to \boxed{g \circ f} \to z$$

Compositions are meant to represent tasks that cannot be carried out in parallel. Imagine that you have two persons working for you, but you can't split the work in half to have them work on it at the same time because the second task cannot be started until the first is finished.

Example 5.2.1: order matters

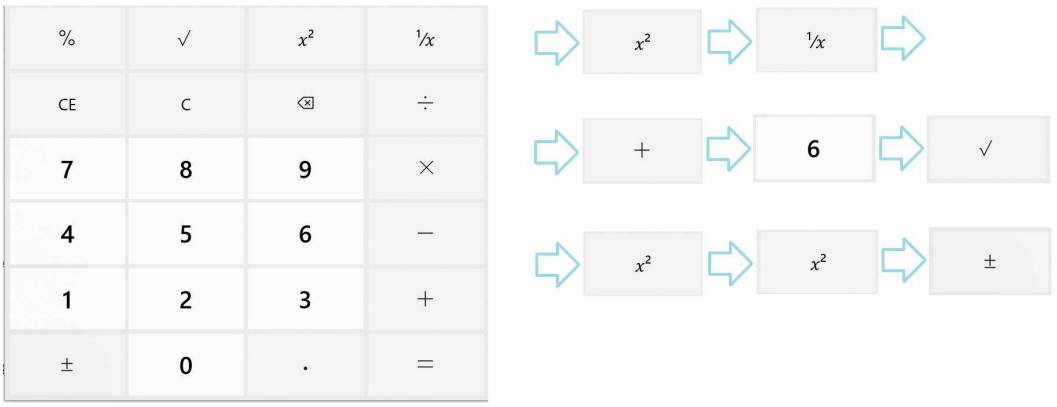
For example, you are making a *chair*. The last two stages are polishing and painting. You can't do them at the same time:

$$\text{chair} \to \boxed{\text{polishing}} \to \boxed{\text{painting}} \to \text{finished chair}$$

You can't change the order either!

Example 5.2.2: computing with calculator

Below, the instructions of the function is "push these buttons" (in that order):



These are the functions created:

$$\frac{1}{x^2}, \quad \sqrt{x+6}, \quad -\left(x^2\right)^2 .$$

Just as with the rest of the algebraic operations, we sometimes want to "undo" compositions. By doing so, we *decompose* the given function into two (or more) simpler parts that can then be addressed separately.

Example 5.2.3: decomposition

Represent $z = h(x) = \sqrt[3]{x^2+1}$ as the composition of two functions.

We need to find an appropriate place to stop in the middle of its computation. One of clues for such a spot might the *radical sign*. What's inside is to be computed first:

$$\sqrt[3]{x^2+1}\,.$$

It's our intermediate variable, say y:

$$y = x^2+1\,.$$

Now to the second stage. We just replace x^2+1 in our formula with y:

$$z = \sqrt[3]{y}\,.$$

This is the decomposition:

$$x \;\mapsto\; x^2+1 = y \;\mapsto\; \sqrt[3]{y} = z\,.$$

To confirm, *substitute* y back into the expression for z.

Example 5.2.4: another decomposition

Decompose:

$$y = \sin(x^2+1)\,.$$

Another clue pointing at the intermediate variable might be *parentheses*:

$$y = \sin\left(x^2+1\right).$$

What's inside will be our new variable:

$$u = x^2+1\,.$$

We substitute that into the original:

$$y = \sin(u)\,.$$

Done! For completeness, we can *name* the functions:

$$u = f(x) = x^2+1,\; y = g(u) = \sin u \implies h(x) = (g\circ f)(x) = \sin(x^2+1)\,.$$

Exercise 5.2.5

Decompose $y = (x+1)^3$.

Exercise 5.2.6

Decompose $y = 2^{x-1}$.

Example 5.2.7: complicated decomposition

Decompose:

$$y = \frac{t^2+1}{t^2-1}.$$

There are no parentheses to use as a clue here. The clue might be the *repetition*. The formula suggests that we can compute once and substitute twice saving computing time:

$$y = \frac{\color{red}{t^2}+1}{\color{red}{t^2}-1}.$$

So, let's try:

$$x = \color{red}{t^2}.$$

We substitute that into the original:

$$y = \frac{\color{red}{x}+1}{\color{red}{x}-1}.$$

Just as valid a choice is

$$u = t^2+1.$$

Then

$$y = \frac{u}{u-2}.$$

What can we say about the choice

$$x = \frac{t^2+1}{t^2-1}?$$

The new function is the same as the original. Such a decomposition is, though technically correct, is pointless as it provides no simplification.

Exercise 5.2.8

Decompose $y = \dfrac{1}{t^2}$.

There may be more than two functions involved in compositions.

Example 5.2.9: gas mileage

Suppose a car is driven at 60 mi/h. Suppose we also know that the car uses 30 mi/gal, while the cost per gallon is \$5. Represent the expense as a function of time.

Consider the flowchart:

$$\begin{array}{ccccccc} \text{time (h)} & \xrightarrow{60\text{ mi/h}} & \text{distance (mi)} & \xrightarrow{30\text{ mi/gal}} & \text{gas used (gal)} & \xrightarrow{5\$/\text{ gal}} & \text{expense (\$)} \\ t & \xrightarrow{f} & d & \xrightarrow{k} & g & \xrightarrow{h} & e \end{array}$$

These are the participating functions:

$$60t = d \qquad \frac{d}{30} = g \qquad 5g = e$$

To answer the question, we substitute from right to left:

$$e = 5g = 5\left(\frac{d}{30}\right) = 5\left(\frac{60t}{30}\right).$$

Simplified:

$$e = 10t\,.$$

Exercise 5.2.10

Redo the example using the functions f, k, h.

Example 5.2.11: order matters

Find $f(g(x))$ and $g(f(x))$ with:

$$f(x) = x^2 \text{ and } g(x) = \cos x\,.$$

The problem may represent a challenge because the variables don't match! We have two problems in one. We seek to recast both problems in terms of these variables:

$$x \;\to\; y \;\to\; z$$

To find $f(g(x))$, first rewrite:

$$f(y) = y^2 \;\text{ and }\; y = g(x) = \cos x\,.$$

Then replace (substitute) y in f with $(\cos x)$, always with parentheses:

$$y^2 \;\longrightarrow\; (\cos x)^2\,,$$

often written as $\cos^2 x$.

To find $g(f(x))$, first rewrite:

$$y = f(x) = x^2 \;\text{ and }\; g(y) = \cos y\,.$$

Then replace y in g with (x^2):

$$\cos y \;\longrightarrow\; \cos(x^2)\,.$$

Example 5.2.12: recursive sequences

Compositions shine a new light on some old ideas. An arithmetic progression is defined as follows:

$n =$	0	1	2	3	4	...
	2,	add 3,	add 3,	add 3,	add 3,	...
$a_n =$	2	5	8	11	14	...

It is just a repeated application of the same function! We have called a sequence a_n *recursive* when its next term is found from the current term by a formula. At its simplest, its next term is found from the current term by applying the same function, which doesn't depend on n:

$$a_{n+1} = F(a_n),\;\; n = 1, 2, 3, ...$$

So, the terms of the sequence are computed – one at a time – by applying the function F repetitively:

$n =$	0	1	2	3	4	...
	a_0,	apply F,	apply F,	apply F,	apply F,	...
$a_n =$	a_0	$a_1 = F(a_0)$	$a_2 = F(a_1)$	$a_3 = F(a_2)$	$a_4 = F(a_3)$	...

It's just one long composition.

Exercise 5.2.13

What is the function that defines a geometric progression? What is different about the factorial?

Example 5.2.14: compositions of functions given by lists

When the two functions are represented by their tables of values, the composition can be computed just as with the rest of algebraic operations (as discussed earlier). It is more complex as we cannot simply go with the same row for all functions. One has to find the right entry in the next function.

Suppose we need to compose these two functions:

t	$x = f(t)$
0	1
1	2
2	3
3	0
4	1

followed by

x	$y = g(x)$
0	5
1	-1
2	2
3	3
4	0

.

The result should be another table of values for the function $h = g \circ f$. To fill this table, we watch where every t goes. What happens to 0 after two steps? We look at the first table: Under f, we have $0 \mapsto 1$. Next, where does 1 go under g? We look at the second table: Under g, we have $1 \mapsto -1$. Therefore, under h, we have $0 \mapsto -1$. That gives us the first row in the new table:

$$f : 0 \mapsto 1 \quad g : 1 \mapsto -1 \quad \Longrightarrow \quad h : 0 \mapsto -1 \,.$$

Furthermore:

$$f : 1 \mapsto 2 \quad g : 2 \mapsto 2 \quad \Longrightarrow \quad h : 1 \mapsto 2 \,.$$

And so on. Some of the computations are shown in this table:

t	x		x	y		
0	1	$\searrow$	0	5		
1	2	$\searrow$	1	-1	$\Longrightarrow$	$0 \to -1$
2	3	$\searrow$	2	2	$\Longrightarrow$	$1 \to 2$
3	0		3	3	$\Longrightarrow$	$2 \to 3$
4	1		4	0		

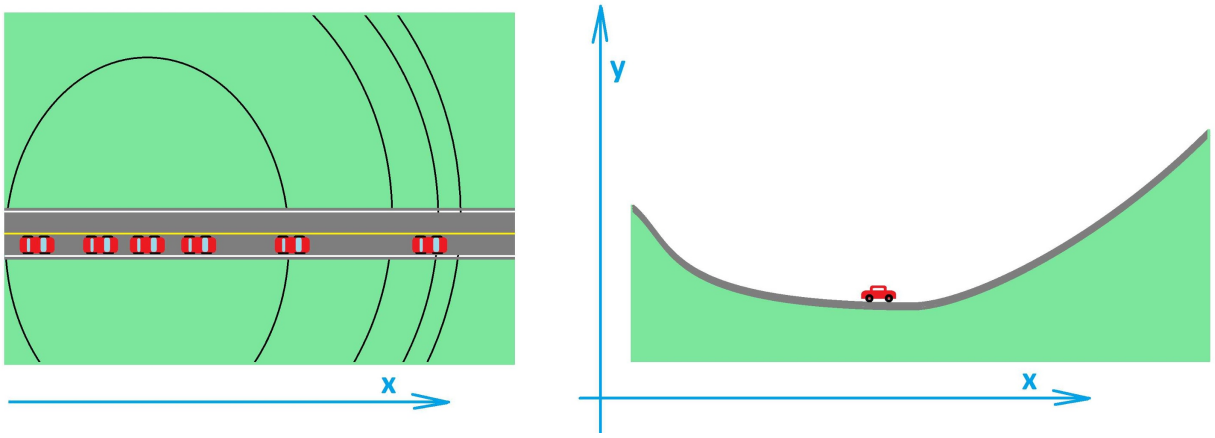
And this is the answer:

$$g \circ f = \begin{array}{|c|c|}\hline t & y = g(f(t)) \\ \hline 0 & -1 \\ 1 & 2 \\ 2 & 3 \\ 3 & 5 \\ 4 & -1 \\ \hline \end{array} \,.$$

Example 5.2.15: driving through terrain

Functions represented by graphs can also be composed. The procedure is, however, more convoluted than those for the rest of the algebraic operations.

Let's review. Suppose a car is driven through a mountain terrain. The location, as seen on a map (left), is known and so is the altitude for each location (right):



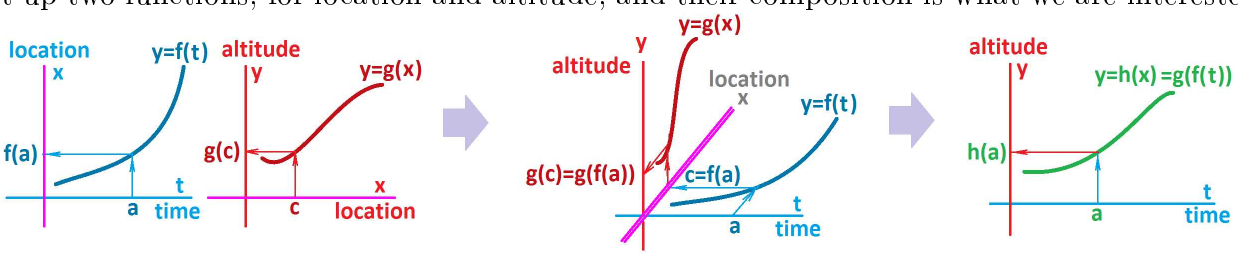
These are the three variables:

- t is time – measured in hr.
- x is the location of the car – measured in mi.
- y is the elevation of the road – measured in ft.

This is their relation:

$$t \;\to\; x \;\to\; y$$

We set up two functions, for location and altitude, and their composition is what we are interested in:



The second function is literally the profile of the road. Here, we have:

- $x = f(t)$ is the location of the car as a function of time.
- $y = g(x)$ is the elevation of the road as a function of (horizontal) location.
- $y = h(t) = g(f(t))$ is the altitude of the road as a function of time.

This is the familiar way to evaluate a function:

$$f(x) = x^2 - x \implies f(3) = 3^2 - 3\,.$$

There is also an alternative notation:

Substitution notation

$$f(x) = x^2 - x \implies f(3) = x^2 - x\Big|_{x=3} = 3^2 - 3$$

This notation also applies to compositions. For example:

- We substitute $z = g(y) = y \cdot 2$ into $u = h(z) = z^2$, which results in the following:

$$u = z^2\Big|_{z=y\cdot 2} = (y \cdot 2)^2\,.$$

- We substitute $y = f(x) = x + 3$ into $z = g(y) = y \cdot 2$, which results in the following:

$$z = y \cdot 2\Big|_{y=x+3} = (x + 3) \cdot 2\,.$$

5.3. Solving equations

Let's review what it means to solve an equation.

We go back to our example of boys and balls. This is our function that tells what ball each boy prefers:

$$\begin{aligned} F(\text{ Tom }) &= \text{basketball} \\ F(\text{ Ned }) &= \text{tennis} \\ F(\text{ Ben }) &= \text{basketball} \\ F(\text{ Ken }) &= \text{football} \\ F(\text{ Sid }) &= \text{football} \end{aligned}$$

So, our function – in the form of this list – answers the question:

▶ Which ball is this boy playing with?

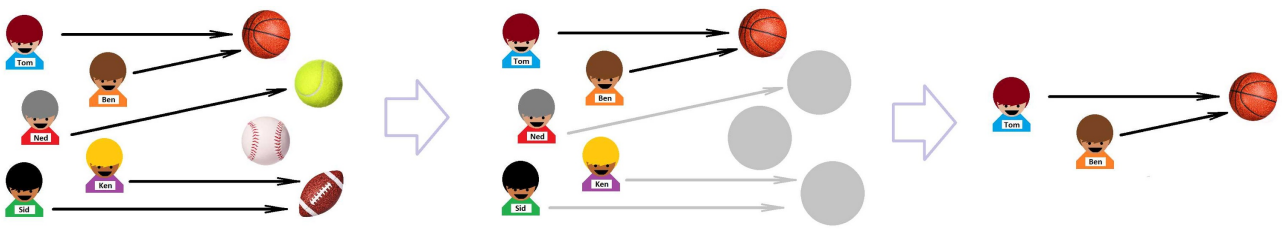
However, what if we turn this question around:

▶ Which boy is playing with this ball?

Let's try an example: Who is playing with the basketball? Before answering it, we can give this question a more compact form, the form of an *equation*:

$$F(\text{ boy }) = \text{ basketball }.$$

Indeed, we need to find the inputs that, under F, produce this specific output. We visualize and answer the question by erasing all irrelevant arrows:



These are a few of possible questions of this kind along with the answers:

- Who is playing with the basketball? Tom and Ben!
- Who is playing with the tennis ball? Ned!
- Who is playing with the baseball? No one!
- Who is playing with the football? Ken and Sid!

It seems that there are several answers to each of these questions. Or are there? "Tom" and "Ben" aren't *two* answers; it's one: "Tom and Ben"! Indeed, if we provide one name and not the other, we haven't fully answered the question. We know that we should write the answer as

$$\{\text{ Tom, Ben }\}.$$

It's a set!

Let's review. The solution of an equation $f(x) = y$ with $f : X \to Y$ is always a set (a subset of X) and it may contain *any* number of elements, including none. To solve an equation with respect to x means to find *all* values of x that satisfy the equation. In other words:

1. When we substitute any of those x's into the equation, we have a true statement.
2. There are no other such x's.

For example, consider how this equation is solved:

$$x+2=5 \implies x=3\,.$$

That's an abbreviated version of the following statement:

▶ If x satisfies the equation $x+2=5$, then x satisfies the equation $x=3$.

Plug in:

$$x+2=(3)+2=5\,.$$

It checks out! We could *try* others and they won't check out:

	$(x)+2=$	$=^{?}5$	TRUE/FALSE	
$x=0$	$(0)+2=2$	$\neq 5$	FALSE	
$x=1$	$(1)+2=3$	$\neq 5$	FALSE	
$x=2$	$(2)+2=4$	$\neq 5$	FALSE	
$x=3$	$(3)+2=5$	$=5$	TRUE	Add it to the list!
$x=4$	$(4)+2=6$	$\neq 5$	FALSE	
...	...		...	

Of course, this trial-and-error method is unfeasible because there are infinitely many possibilities.

A method of *how* we may arrive to the answer is discussed in this section.

Recall the basic methods ("rules") of *handling* equations.

Example 5.3.1: simple equations

In order to solve the equation

$$x+2=5\,,$$

subtract 2 from both sides producing

$$x+2-2=5-2 \implies x=3\,.$$

In order to solve the equation

$$3x=2\,,$$

divide by 3 both sides producing

$$3x/3=2/3 \implies x=2/3\,.$$

This is the summary.

Theorem 5.3.2: Basic Algebra of Equations

- *Multiplying both sides of an equation by a number preserves it. In other words, we have:*
$$a=b \implies ka=ka \quad \text{for any } k\,.$$
- *Adding any number to both sides of an equation preserves it. In other words, we have:*
$$a=b \implies a+s=b+s \quad \text{for any } s\,.$$

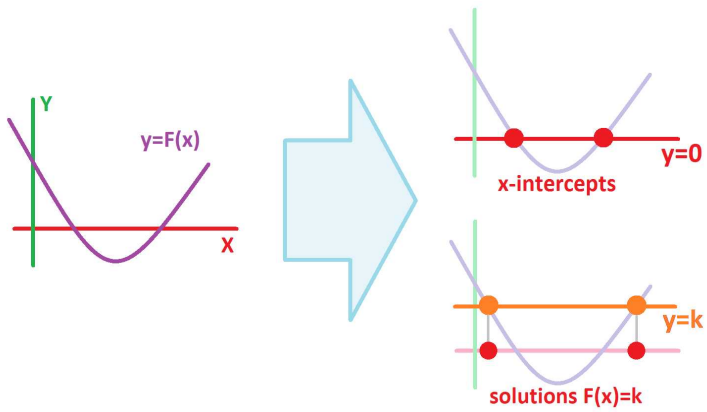
Warning!

Multiplying an equation by 0 is pointless.

Exercise 5.3.3

What about the converses?

As a reminder, in the special case when the right-hand side of the equation is zero, the solution set to this equation has a clear *geometric* meaning: An *x-intercept* of a numerical function f is any solution to the equation $f(x) = 0$. In other words, these are the x-coordinates of the *intersections* of the graph with the x-axis (top):

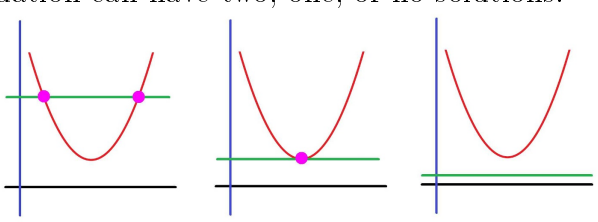


Furthermore, when the right-hand side is a number, say, k, the solution to this equation $f(x) = k$ gives us the intersection of the graph with the line $y = k$ (bottom).

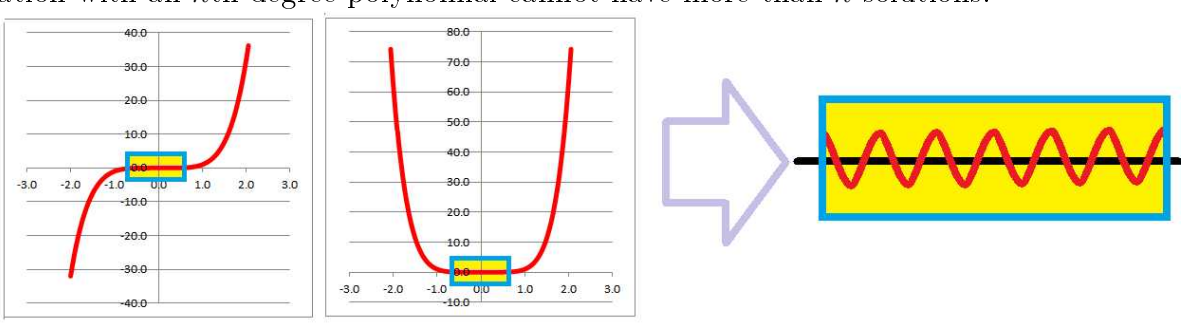
Example 5.3.4: counting solutions

Our knowledge of the shape of the graph of a particular function will tell us how many solutions such an equation might have.

For example, a quadratic equation can have two, one, or no solutions:

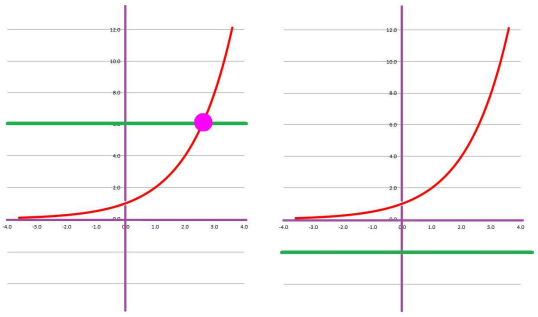


An equation with an nth degree polynomial cannot have more than n solutions:

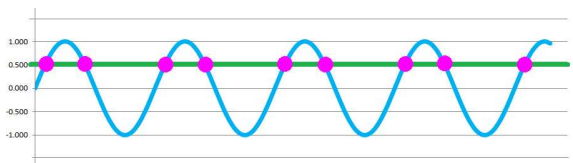


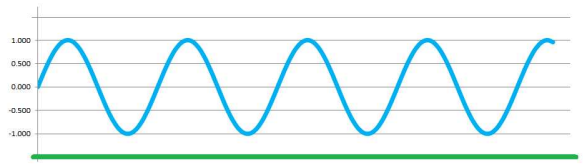
But an equation with an odd degree polynomial will have at least one solution!

An equation with the exponential function can have one solution or none:



An equation with a periodic function (such as the sine) can have infinitely many solutions or none:





Our interest in this section, though, is *algebra*.

We will deal with a simple kind of equation:

- x is present only once (in the left-hand side).

Like this:

$$x^2 = 17\,.$$

Starting with such an equation, our goal is – through a series of manipulations – to arrive to an even simpler kind of equation:

- x is *isolated* (in the left-hand side).

Like this:

$$x = \sqrt{17}\,.$$

Warning!

This *is* an equation.

In other words, we will try to find ways to get from left to right here:

$$\sin\left(e^{\frac{1}{\sqrt{x}}}\right) = 5 \quad \longrightarrow \quad ??? \quad \longrightarrow \quad x = \underbrace{\cdots\cdots}_{\text{no } x \text{ here}}$$

The *main idea* of how to manipulate equations is as follows:

- We apply a function to both sides of the equation, producing a new equation.

For example, if we have an equation, say,

$$x + 2 = 5\,,$$

we treat it as a number, call it y. Then we deal with this number:

$$y = x + 2 = 5, \quad \text{apply } z = y - 2 \implies (x+2) - 2 = 3 - 2 \implies x = 3 \quad \text{Solved!}$$

The idea is to produce – from an equation satisfied by x – another equation satisfied by x.

However, the challenge (and an opportunity) is that applying *any* function in this manner will produce a new equation satisfied by x! For example:

$$\begin{array}{llll} y = x + 2 = 5, & \text{apply } z = y + 2 & \implies (x+2) + 2 = 5 + 2 \implies x + 4 = 7 & \text{Not solved!} \\ y = x + 2 = 5, & \text{apply } z = y^2 & \implies (x+2)^2 = 5^2 & \text{Not solved!} \\ y = x + 2 = 5, & \text{apply } z = \sin y & \implies \sin(x+2) = \sin 5 & \text{Not solved!} \end{array}$$

Indeed, it is the *definition* of a function that every input has exactly one output. Suppose we have a function g. Then, two equal (according to the equation) inputs of g will produce two equal outputs (another equation), always:

$$\begin{array}{rccc} \text{old equation:} & a & = & b \\ & \downarrow g & & \downarrow g \\ \text{new equation:} & g(a) & = & g(b) \end{array}$$

In particular, we have, for any function g:

$$x+2=5 \implies g(x+2)=g(5)\,.$$

That is why if the first equation is satisfied by x, then so is the second.

There are infinitely many possibilities for this function g:

$$\begin{array}{ccccc} x+4=x+2+2=7 & & (x+2)^2=5^2 & & \sin(x+2)=\sin 5 \\ & \nwarrow & \uparrow & \nearrow & \\ x+5=x+2+3=8 & \leftarrow & \boxed{x+2=5} & \rightarrow & 2^{x+2}=2^5 \\ & \swarrow & \downarrow & \searrow & \\ \boxed{x=x+2-2=3} & & (x+2)^3=5^3 & & \sqrt{x+2}=\sqrt{5} \end{array}$$

If x satisfies the equation in the middle, it also satisfies the rest of the equations.

If we want to solve the original equation, which function – out of infinitely many – do we pick? Some of them clearly make the equation *more* complex than the original! It is the challenge for the equation solver to have enough foresight to choose a function to apply that will make the equation *simpler*.

Example 5.3.5: solving equation with function given by flowchart

Let's consider this equation:

$$3\cdot\left(\frac{\sqrt{x}}{4}+1\right)=6\,.$$

Here, several functions are consecutively applied to x. This is the flowchart of this function:

$$f:\ x \rightarrow \boxed{\text{root}} \rightarrow \boxed{\text{divide by 4}} \rightarrow \boxed{\text{add 1}} \rightarrow \boxed{\text{multiply by 3}} \rightarrow y$$

We can plug in any value on the left and get the output on the right:

$$0 \rightarrow \boxed{\text{root}} \rightarrow 0 \rightarrow \boxed{\text{divide by 4}} \rightarrow 0 \rightarrow \boxed{\text{add 1}} \rightarrow 1 \rightarrow \boxed{\text{multiply by 3}} \rightarrow 3$$

We *test* possible inputs this way. That one has failed; it's not 6!

Is there a better method?

We would like to get to x. To get to it, we will need to undo these functions one by one. In what order? Right to left, of course. We reverse the flow of the flowchart:

$$\begin{array}{lccccccccccc} f: & x & \rightarrow & \boxed{\text{root}} & \rightarrow & \boxed{\text{divide by 4}} & \rightarrow & \boxed{\text{add 1}} & \rightarrow & \boxed{\text{multiply by 3}} & \rightarrow & 6 \\ f^{-1}: & x & \leftarrow & \boxed{\text{square}} & \leftarrow & \boxed{\text{multiply by 4}} & \leftarrow & \boxed{\text{subtract 1}} & \leftarrow & \boxed{\text{divide by 3}} & \leftarrow & 6 \end{array}$$

Of course, the pairs of functions aligned vertically are the *inverses* of each other:

$$\begin{array}{lccccccccccc} x & \rightarrow & \boxed{\text{root}} & \rightarrow & \boxed{\text{divide by 4}} & \rightarrow & \boxed{\text{add 1}} & \rightarrow & \boxed{\text{multiply by 3}} & \rightarrow & y \\ & & \text{inverse?} & & \text{inverse} & & \text{inverse} & & \text{inverse} & & \downarrow \\ x & \leftarrow & \boxed{\text{square}} & \leftarrow & \boxed{\text{multiply by 4}} & \leftarrow & \boxed{\text{subtract 1}} & \leftarrow & \boxed{\text{divide by 3}} & \leftarrow & y \end{array}$$

They, therefore, can be canceled out one pair at a time:

$$\begin{array}{lccccccc} x & \rightarrow & \boxed{\text{root}} & \rightarrow & \boxed{\text{divide by 4}} & \rightarrow & \boxed{\text{add 1}} & \boxed{\blacksquare\blacksquare\blacksquare} \\ & & & & & & \downarrow & \\ x & \leftarrow & \boxed{\text{square}} & \leftarrow & \boxed{\text{multiply by 4}} & \leftarrow & \boxed{\text{subtract 1}} & \boxed{\blacksquare\blacksquare\blacksquare} \end{array}$$

Then, we have the following and we can cancel more:

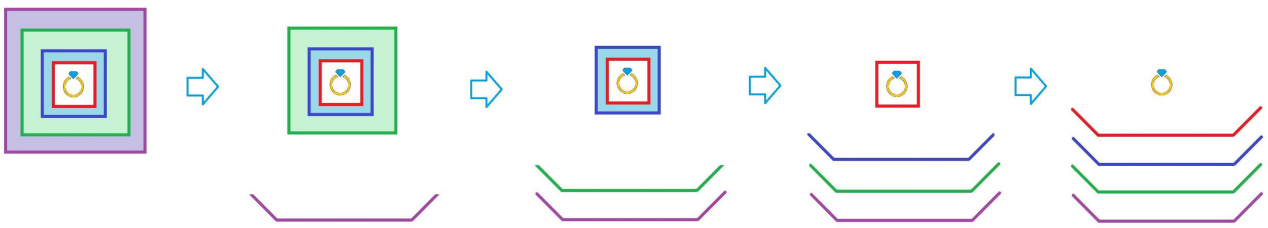
$$x \;\to\; \boxed{\text{root}} \;\to\; \boxed{\text{divide by 4}} \;\to\; \boxed{\blacksquare\blacksquare\blacksquare} \;\to\; \boxed{\blacksquare\blacksquare\blacksquare}$$
$$\downarrow$$
$$x \;\leftarrow\; \boxed{\text{square}} \;\leftarrow\; \boxed{\text{multiply by 4}} \;\leftarrow\; \boxed{\blacksquare\blacksquare\blacksquare} \;\leftarrow\; \boxed{\blacksquare\blacksquare\blacksquare}$$

And so on. We have demonstrated that the second row is indeed the inverse of the first.

With this fact understood, we find x by starting on the left with 6:

$$\pm 16 \leftarrow \boxed{\text{square}} \leftarrow 4 \leftarrow \boxed{\text{multiply by 4}} \leftarrow 1 \leftarrow \boxed{\text{subtract 1}} \leftarrow 2 \leftarrow \boxed{\text{divide by 3}} \leftarrow 6$$

Such equations can be visualized as follows; we see a single x "wrapped" in several layers of functions, as if a gift:



To get to the gift, the only method is to remove one wrapper at a time, from the outside in. In fact, you don't even know what kind of wrapper is the next until you've removed the last one!

Warning!

As we are unwrapping the left-hand side, we are *wrapping* the right-hand side in the layers of the inverses of these functions; we are only creating more work if there is an x there, as in:

$$3 \cdot \left(\frac{\sqrt{x}}{4} + 1 \right) = x \,.$$

Example 5.3.6: unwrapping layers of functions

What if we have *several* functions applied consecutively to x? Which function do we choose to apply?

Consider:

$$\text{Equation 1:} \quad 5 \cdot \left(\left(\frac{x}{2} + 1 \right)^2 + 3 \right) - 17 = 3$$

The last operation on the left is -17. That's the function we face, and the inverse of this function is to be applied. We choose, therefore:

$$f(z) = z{-}17 \,,$$

where

$$z = 5 \cdot \left(\left(\frac{x}{2} + 1 \right)^2 + 3 \right) .$$

Then we apply

$$g(y) = f^{-1}(y) = y{+}17 \,.$$

We conclude:

$$z{-}17 = 3 \implies (z{-}17){+}17 = 3{+}17 \implies z = 20 \,.$$

We have a new equation now:

$$\text{Equation 2:} \quad 5 \cdot \left(\left(\frac{x}{2} + 1 \right)^2 + 3 \right) = 20$$

We pause to appreciate the fact that we face a *simpler* equation than the original!

We continue in this fashion. As we progress, we apply the inverse of the function that appears first in the equation, i.e., it is applied last. We have the following sequence of steps "unwrapping" the functions one at a time:

$$\begin{aligned}
&\text{Equation 2:} && 5\cdot\left(\left(\frac{x}{2}+1\right)^2+3\right)=20 && \implies && 5\cdot\left(\left(\frac{x}{2}+1\right)^2+3\right)/5=20/5 && \implies\\
&\text{Equation 3:} && \left(\frac{x}{2}+1\right)^2+3=4 && \implies && \left(\frac{x}{2}+1\right)^2+3-3=4-3 && \implies\\
&\text{Equation 4:} && \left(\frac{x}{2}+1\right)^2=1 && \implies && \sqrt{\left(\frac{x}{2}+1\right)^2}=\sqrt{1} && \overset{?}{\implies}\\
&\text{Equation 5:} && \frac{x}{2}+1=1 && \implies && \frac{x}{2}+1-1=1-1 && \implies\\
&\text{Equation 6:} && \frac{x}{2}=0 && \implies && \frac{x}{2}\cdot 2=0\cdot 2 && \implies\\
&\text{Equation 7:} && x=0\,. &&
\end{aligned}$$

We are finished!

But wait a minute, $x=-4$ is also a solution! What happened? How did we lose it? We forgot that there are *two* inverses of $y=x^2$ and disregarded the latter of these two cases:

$$\begin{aligned}
&\text{case 1:} && z\geq 0 && \implies && \sqrt{z^2}=z\,.\\
&\text{case 2:} && z\leq 0 && \implies && \sqrt{z^2}=-z\,.
\end{aligned}$$

We need to re-do the solution starting at the question mark. The equation produces two cases (the latter is finished above):

$$\sqrt{\left(\frac{x}{2}+1\right)^2}=\sqrt{1}$$

$\swarrow$ $\qquad\qquad$ $\searrow$

$z=\frac{x}{2}+1\geq 0$			OR	$z=\frac{x}{2}+1\leq 0$		
$\left(\frac{x}{2}+1\right)$	$=1$	$\implies$	OR	$-\left(\frac{x}{2}+1\right)$	$=1$	$\implies$
$\frac{x}{2}+1$	$=1$	$\implies$	OR	$\frac{x}{2}+1$	$=-1$	$\implies$
$\frac{x}{2}$	$=0$	$\implies$	OR	$\frac{x}{2}$	$=-2$	$\implies$
x	$=0$		OR	x	$=-4$	

Both $x=1$ and $x=-4$ satisfy their respective conditions. Therefore, the solution set is the following:

$$\{x:\ x=0\quad\texttt{OR}\quad x=-4\}=\{0\}\cup\{-4\}=\{0,-4\}\,.$$

Exercise 5.3.7

Solve the equation:

$$5\cdot\left(\left(\frac{x^2}{2}+1\right)^2+3\right)-17=3\,.$$

Exercise 5.3.8

Solve the equation:

$$\left(5\cdot\left(\left(\frac{x}{2}+1\right)^2+3\right)-17\right)^3=27\,.$$

Make up your own equation and solve it. Repeat.

Example 5.3.9: not one-to-one?

If between us and x there is a function, we would like to remove it. How? Based on our experience, the answer is: Apply the *inverse* of the function that we face! We did that above:

$$x\mapsto f(x)=x+2\implies y\mapsto g(y)=f^{-1}(y)=y-2\,.$$

Is applying the inverse of the function a *fool-proof plan*? No. Try this:

$$x^2=1,\text{ apply }z=\sqrt{y}\implies\sqrt{x^2}=\sqrt{1}\implies x=1\ ??$$

We have lost $x=-1$! It satisfies the first equation but not the last. What happened? There is nothing wrong with the logic of applying the inverse and producing a new equation satisfied by x; however, the cancellation of the function and its "inverse" was incorrect. The square and the square root are inverses of each other (and cancel) only, separately, on the rays $[0,+\infty)$ and $(-\infty,0]$. We disregarded the latter of these two cases:

$$\text{case 1: }\ x\geq 0\implies\sqrt{x^2}=x,\quad\text{case 2: }\ x\leq 0\implies\sqrt{x^2}=-x\,.$$

We can point out exactly why this has failed; the function isn't *one-to-one*!

Example 5.3.10: not onto?

Is applying the inverse of the function – making sure that it is one-to-one – a *fool-proof plan*? No. Try this:

$$\sqrt{x}=-1,\text{ apply }z=y^2\implies\left(\sqrt{x}\right)^2=(-1)^2\implies x=1\ ??$$

We have a solution where there is none! It satisfies the last equation but not the first. What happened? There is nothing wrong with our logic; however, the cancellation of the function and its "inverse" was incorrect, once again. The square and the square root are inverses of each other (and cancel) only, separately, on the rays $[0,+\infty)$ and $(-\infty,0]$. We can point out exactly why this failed; the function isn't *onto*!

Example 5.3.11: solving equation

Consider the equation:

$$\sqrt{x+1}=3\,.$$

Examining the equation, we see x on the left only and it is subjected to two functions, the last of which is the square root. Therefore, to make a step toward isolating x, square both sides:

$$\left(\sqrt{x+1}\right)^2=3^2\,.$$

If we were to cancel the two functions on the left as inverses, we get this new equation:

$$x+1=9\,.$$

The solution is $x=8$. But was the cancellation valid? Instead of considering the function, we just confirm that the solution falls within the domain, which is $x+1\geq 0$, of the original equation. It does.

Exercise 5.3.12

Solve the equation:

$$\sqrt{x^2+1} = 3\,.$$

Exercise 5.3.13

Solve the equation:

$$5\sqrt{x+1} = 3\,.$$

Make up your own equation and solve it. Repeat.

Using only invertible functions ensures that we don't lose solutions or gain non-solutions. This is why this approach will often allow us to enhance our solution method: from

- "`IF` x satisfies an equation, `THEN` it satisfies the next" to
- "x satisfies an equation `IF AND ONLY IF` it satisfies the next".

They are equivalent!

The advantage of this approach is that the equations will have *the same solution set* as we progress through the stages. For example, this is how we would rather present the solution of the very first equation in this section:

$$x+2 = 5 \iff x = 3\,.$$

That's an abbreviated version of the following statement:

▶ x satisfies $x+2=5$ `IF AND ONLY IF` x satisfies $x=3$.

Repeated as many times as necessary, the solution set of each equation is the same as that of the original equation!

Example 5.3.14: solving equations

Two complete solutions are shown below:

$$\begin{array}{ll} (1) & 2(x+2)-3 = 5x \iff 2x+4-3 = 5x \iff -3x = -1 \iff x = 1/3\,. \\ (2) & x^2+1 = 0 \iff x^2 = -1 \iff \varnothing\,. \end{array}$$

So, to solve the type of equation we face – the variable x subjected to a sequence of functions – we apply the inverses of these functions in the reversed order. We may have to split the domains of the functions so that they are both one-to-one and onto. As we have seen, this step also splits our equation.

Example 5.3.15: split domain

A solution is shown below:

$$x^2 = 1 \iff x = -1 \quad \texttt{OR} \quad x = 1\,.$$

Exercise 5.3.16

Solve the equation in this manner

$$\sqrt{x+1} = 3\,.$$

Exercise 5.3.17

Solve the equation in this manner

$$\sqrt{x^2+1} = 3\,.$$

Make up your own equation and solve it. Repeat.

Example 5.3.18: split domain

What if, instead of a single x, we face an expression that depends on x? For example, we might face this:

$$5 \cdot \Big(\big(x^2 + x + 1\big)^2 + 3 \Big) - 17 = 3\,.$$

But what is an "expression" if not just another variable, $y = x^2 + x + 1$, represented as a function of x. To make this choice clear, we substitute:

$$5 \cdot \Big((y)^2 + 3 \Big) - 17 = 3\,.$$

In that we case, we go after this variable, y, by following the same plan as before: remove these layers one by one. Once y is found, as in the example):

$$y = 0 \quad \texttt{OR} \quad y = -4\,.$$

we have a much simpler equation, or a combination of equations, for x:

$$x^2 + x + 1 = 0 \quad \texttt{OR} \quad x^2 + x + 1 = -4\,.$$

Exercise 5.3.19

Finish the solution. Make up your own equation and solve it. Repeat.

As a summary, we have developed the following method of solving equations.

Theorem 5.3.20: General Algebra of Equations

Suppose g is an invertible function. Then applying g to both sides of an equation creates an equivalent equation. In other words, we have:

$$a = b \iff g(a) = g(a)\,.$$

Furthermore, the two equations have the same solution set.

Exercise 5.3.21

What is the relation between the two solution sets if instead of $\iff$, we have (a) $\Longleftarrow$, (b) $\Longrightarrow$?

5.4. The algebra of logarithms

Let's go all the way back to the definition of the exponent as repeated multiplication (Chapter 1):

$$\underbrace{a \cdot a \cdot \ldots \cdot a}_{x \text{ times}} = a^x\,,$$

for any $x = 1, 2, 3, \ldots$. We defined the logarithm base a of y to be "the power I am to raise a to get y". When x is a natural number, this simply means "the number of times I multiply a by itself to get y":

$$\log_a \big(\underbrace{a \cdot a \cdot \ldots \cdot a}_{x \text{ times}} \big) = x\,.$$

From the *Rules of Exponents* theorem, we will derive the corresponding rules about the logarithm.

We will continue to use the *Cancellation Laws*:

$$a^{\log_a y} = y \quad \Rightarrow \quad a^{\log_a y} = y \quad \Rightarrow \quad y = y$$

$$\log_a a^x = x \quad \Rightarrow \quad \log_a a^x = x \quad \Rightarrow \quad x = x$$

Before we consider the major rules, let's establish two very basic *values* of the logarithm:

$$a^0 = 1 \iff \log_a 1 = 0$$

and

$$a^1 = a \iff \log_a a = 1$$

We turn to the *Addition-Multiplication Rule* (#1) now:

$$a^{x+y} = a^x \cdot a^y \,.$$

In the case of natural x and y, it comes from counting two strings of a's:

$$a^x \cdot a^y = \overbrace{\underbrace{a \cdot a \cdot \ldots \cdot a}_{x \text{ times}} \cdot \underbrace{a \cdot a \cdot \ldots \cdot a}_{y \text{ times}}}^{x+y \text{ times}} = \overbrace{\underbrace{a \cdot a \cdot \ldots \cdot a}_{a^x} \cdot \underbrace{a \cdot a \cdot \ldots \cdot a}_{a^y}}^{a^{x+y}} = a^{x+y} \,.$$

So, we can see that when two numbers are multiplied, the powers, i.e., the logarithms, are *added*:

$$\log_a \left(a^x \cdot a^y\right) = \log_a \left(\overbrace{\underbrace{a \cdot a \cdot \ldots \cdot a}_{x \text{ times}} \cdot \underbrace{a \cdot a \cdot \ldots \cdot a}_{y \text{ times}}}^{x+y \text{ times}} \right) = x + y \,.$$

Let's apply $\log_a$ to this equation and then concentrate on $X = a^x$ and $Y = a^y$.

Theorem 5.4.1: Addition-Multiplication Rule of Logarithms

For any $a > 0$ and any $X, Y > 0$, we have:

$$\log_a(X \cdot Y) = \log_a X + \log_a Y$$

Proof.

We aim to apply *Addition-Multiplication Rule of Exponents*. But what are those exponents? We define:

$$x = \log_a X, \; y = \log_a Y \iff X = a^x, \; Y = a^y \,.$$

We now apply the logarithm to both sides of the rule:

$$\begin{array}{rrcll} & a^{x+y} & = & a^x \cdot a^y & \text{Apply } \log_a \,. \\ \iff & \log_a(a^{x+y}) & = & \log_a(a^x \cdot a^y) & \text{Now cancel.} \\ \iff & x + y & = & \log_a(a^x \cdot a^y) & \text{Now substitute.} \\ \iff & \log_a X + \log_a Y & = & \log_a(X \cdot Y) \,. & \end{array}$$

Example 5.4.2: expand-contract

In an attempt to "simplify" an expression, we might have to carry out the following *opposite* tasks.

First, we may have to *expand* a logarithm into several by reading the last formula from left to right:

$$\log_2(24) = \log_2(2^3 \cdot 3) = \log_2(2^3) + \log_2(3) = 3 + \log_2 3\,.$$

This requires factoring the number inside the logarithm. Having one of the factors equal to the base helps to simplify further.

Second, we may have to *contract* several logarithms into one by reading the formula from right to left:

$$\log_3(7) + \log_3(2) + 1 = \log_3(7 \cdot 2) + \log_3(3) = \log_3(7 \cdot 2 \cdot 3) = \log_3(56)\,.$$

The need for one or the other depends on the context or the goal we are pursuing.

This is identical to what we did in the past with the *Binomial Formula*:

$$\begin{array}{lll} (x+1)^2 & = x^2 + 2x + 1, & \text{expand} \\ \text{contract,} & 9x^2 + 6x + 1 & = (3x+1)^2 \end{array}$$

Exercise 5.4.3

Expand the logarithm $\log_3(24)$. Make your own expression and expand it. Repeat.

Exercise 5.4.4

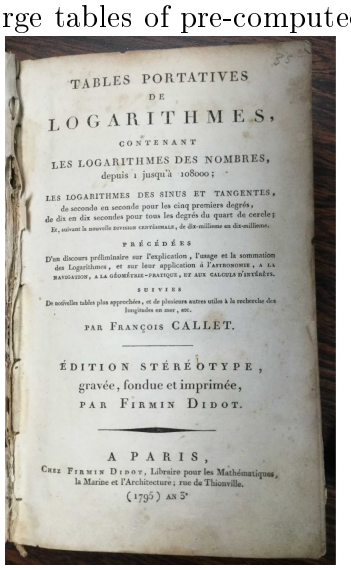
Contract this to a single logarithm: $\log_2(27) + \log_2(2) + 0$. Make your own expression and contract it. Repeat.

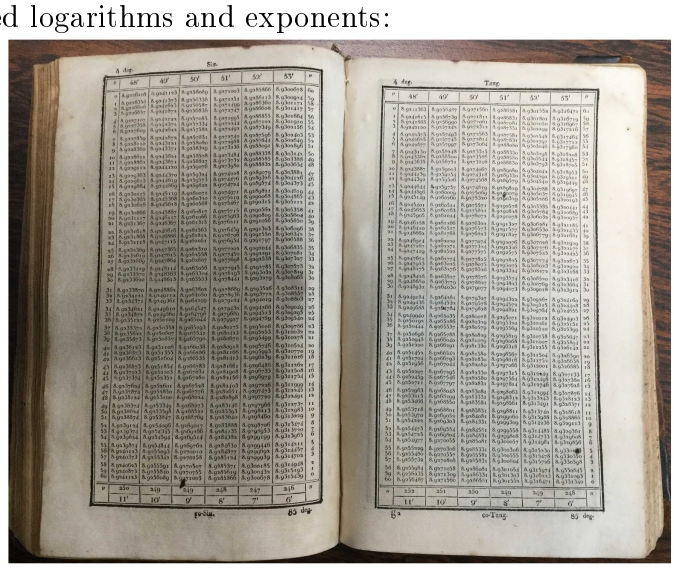
Example 5.4.5: tables of logarithms

The logarithms were used for computations in the times before computers. Specifically, the Addition-Multiplication Rule of Exponents was used to *multiply large numbers*. The key idea is that it converts multiplication to addition, which is easier:

$$\begin{array}{rcccl} & \log_a\big(X & \underbrace{\times}_{\text{multiplication}} & Y\big) \\ = & \log_a X & \underbrace{+}_{\text{addition}} & \log_a Y \end{array}$$

They used large tables of pre-computed logarithms and exponents:


TABLES PORTATIVES
DE
LOGARITHMES,
CONTENANT
LES LOGARITHMES DES NOMBRES,
depuis 1 jusqu'à 108000;
LES LOGARITHMES DES SINUS ET TANGENTES,
de seconde en seconde pour les cinq premiers degrés,
de dix en dix secondes pour tous les degrés du quart de cercle;
Et, suivant la nouvelle division centésimale, de dix-millième en dix-millième.
PRÉCÉDÉES
D'un discours préliminaire sur l'explication, l'usage et la sommation des Logarithmes, et sur leur application à l'Astronomie, à la Navigation, à la Géométrie-pratique, et aux calculs d'intérêts.
SUIVIES
De nouvelles tables plus approchées, et de plusieurs autres utiles à la recherche des longitudes en mer, etc.
PAR François CALLET.
ÉDITION STÉRÉOTYPE,
gravée, fondue et imprimée,
PAR Firmin Didot.
A PARIS,
Chez Firmin Didot, Libraire pour les Mathématiques, la Marine et l'Architecture; rue de Thionville.
(1795) an 5e



To multiply X by Y, follow these steps:

$$\begin{array}{lll} X \to & \boxed{\text{table}} & \to \log_a X = x \\ & & \searrow \\ & & x+y=z \to \boxed{\text{table}} \to a^z = X \cdot Y, \quad \text{done.} \\ & & \nearrow \\ Y \to & \boxed{\text{table}} & \to \log_a Y = y \end{array}$$

The computation amounts to using the tables three times and doing *addition* (instead of multiplication) once. For example, to multiply 1234 and 4321 we follow these steps:

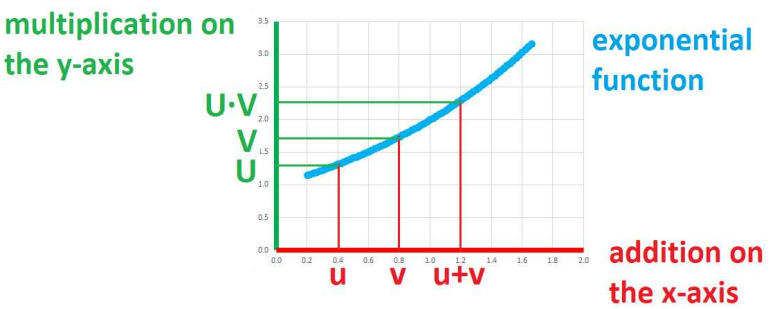
$$\begin{array}{rrrll} 1234 & \to \boxed{\text{table}} \to & \ln 1234 & = 7.118016204 \\ & & & + \\ 4321 & \to \boxed{\text{table}} \to & \ln 4321 & = 8.371242136 \\ \hline & & & = 15.48925834 \\ & \to \boxed{\text{table}} \to & e^{15.48925834} & = 5332114 \end{array}$$

The same method of addition is used in the design of the slide rule.

The two rules – of exponents and logarithms – match:

$$\begin{array}{rcl} a^{x+y} & = & a^x \cdot a^y, \\ \log_a X + \log_a Y & = & \log_a (X \cdot Y) \end{array}$$

In fact, these addition signs on the left and the multiplication signs on the right match up too! So, the addition that is happening in the x-axis is transformed by the exponential function to the y-axis where it becomes multiplication; the logarithm does the opposite:



Next, if

- the logarithms turn multiplication into addition, then
- the logarithms turn division into subtraction.

Indeed, division is just multiplication by a negative power. Therefore, we have the following:

$$\begin{array}{rcl} a^{x-y} & = & a^x \div a^y \\ \log_a X - \log_a Y & = & \log_a (X \div Y) \end{array}$$

Here is the simplest instance of this rule:

$$\log_a \frac{1}{Y} = -\log_a Y$$

Example 5.4.6: expand and contract

These formulas can be read from left to right:

$$\log_3 6 - \log_3 18 = \log_3 \frac{6}{18} = \log_3 \frac{1}{3} = \log_3 3^{-1} = -1\,;$$

or from right to left:

$$\log_5 \frac{1}{25} = \log_5 1 - \log_5 25 = -2\,.$$

Exercise 5.4.7

Expand the logarithm $\log_2(.125)$. Make your own expression and expand it. Repeat.

Exercise 5.4.8

Contract this to a single logarithm: $\log_2(27) - \log_2(2) + 1$. Make your own expression and contract it. Repeat.

Now the Multiplication-Exponentiation Rule (#3):

$$a^{xy} = (a^x)^y\,.$$

In the case of natural x and y, it comes from counting repeated strings of a's:

$$\left(a^x\right)^y = \left.\begin{matrix} a^x \\ a^x \\ ,\ldots \\ a^x \end{matrix}\right\}\ y \text{ times} = \left.\begin{matrix} \underbrace{a\cdot a\cdot \ldots \cdot a}_{x \text{ times}} \\ \underbrace{a\cdot a\cdot \ldots \cdot a}_{x \text{ times}} \\ \ldots \\ \underbrace{a\cdot a\cdot \ldots \cdot a}_{x \text{ times}} \end{matrix}\right\}\ y \text{ times} = \underbrace{a\cdot a\cdot \ldots \cdot a}_{xy \text{ times}} = a^{xy}\,.$$

So, we can see that when a power is taken to another power, the former power, i.e., the logarithm, is *multiplied* by the latter power:

$$\log_a\left(\left(a^x\right)^y\right) = \log_a\left(\underbrace{a^x\cdot a^x\cdot \ldots \cdot a^x}_{y \text{ times}}\right) = \log_a\left(\underbrace{a\cdot a\cdot \ldots \cdot a}_{xy \text{ times}}\right) = xy\,.$$

Let's concentrate on $X = a^x$.

Theorem 5.4.9: Multiplication-Exponentiation Rule of Logarithms

For any $a > 0$ and any $x, y > 0$, we have:

$$\log_a(X^y) = y \cdot \log_a X$$

Proof.

We aim to apply the *Multiplication-Exponentiation Rule of Exponents*. But what is this exponent? We define:

$$x = \log_a X \iff X = a^x\,.$$

We now apply the logarithm to both sides of the identity:

$$\begin{aligned} a^{xy} &= (a^x)^y && \text{Apply } \log_a . \\ \iff \log_a(a^{x\cdot y}) &= \log_a\left((a^x)^y\right) && \text{Now cancel.} \\ \iff x\cdot y &= \log_a\left((a^x)^y\right) && \text{Now substitute.} \\ \iff \log_a X\cdot y &= \log_a(X^y)\,. \end{aligned}$$

The formula answers the question about what happens if we apply a logarithm to an exponential function with a *different* base. Though not a complete cancellation (as in the case of equal bases), it turns the composition into a *linear function*:

$$\log_a(b^y) = \log_a b\cdot y\,.$$

Of course, choosing $b = a$ brings us back to the *Cancellation Law*, i.e., the definition:

$$\log_a(a^y) = y\log_a a = y\,.$$

Example 5.4.10: simplify

In an attempt to "simplify" an expression, we might have to carry out the following *opposite* tasks.

First, we may have to *expand* a logarithm into a multiple of a simpler logarithm by reading the formula from left to right:

$$\log_5(32) = \log_5(2^5) = 5\log_5 2\,.$$

Second, we may have to *contract* a multiple of a logarithm into single one by reading the formula from right to left:

$$\frac{1}{2}\log_3 7 = \log_3(7^{1/2}) = \log_3\sqrt{7}\,.$$

The need for one or the other depends on the context or the goal we are pursuing.

The rule says that the exponent inside the logarithm travels to the outside and becomes a factor:

$$\log_a\left(X^{\textcircled{y}}\right) = \textcircled{y}\log_a X$$

Exercise 5.4.11

Expand the logarithm $\log_2(100)$. Make your own expression and expand it. Repeat.

Exercise 5.4.12

Contract this to a single logarithm: $\frac{1}{3}\log_2(125)$. Make your own expression and contract it. Repeat.

Let's consider again the *exponential models* (Chapter 4):

$$f(x) = Ca^x\,.$$

Here x is time and $f(x)$ is a quantity that grows or declines with time, in multiples of a.

With the help of the logarithm, we are able this time to solve a lot more problems about these models. The nature of these problems is "inverse" to those we considered previously. Below we list some familiar models

along with the old and the new problems:

Models	**Direct problems: substitute**	**Inverse problems: solve equation**
1. Compounded interest: \$1000 deposit, 10% APR. $f(x) = 1000 \cdot 1.1^x$	Q: How much after 10 years? Substitute: $f(10) = 1000 \cdot 1.1^{10}$ $\approx \$2,593$.	Q: How long to reach \$2000? Solve: $1000 \cdot 1.1^x = 2000 \Rightarrow 1.1^x = 2$ $\Rightarrow x = \log_{1.1} 2 \approx 7.27$ years.
2. Population decline: 1M current, 5% decline. $f(x) = 1 \cdot .95^x$	Q: How large after 5 years? Substitute: $f(5) = .95^5$ $\approx 774K$.	Q: How long to reach 500K? Solve: $1 \cdot .95^x = .5 \Rightarrow .95^x = .5$ $\Rightarrow x = \log_{.95} .5 \approx 13.51$ years.
3. Radiocarbon decay: 5000 years half-life. $f(x) = .5^{x/5000}$	Q: How much left after 1000 years? Substitute: $f(5) = .5^{1000/5000}$ $\approx 87\%$.	Q: How long to lose 90%? Solve: $.5^{x/5000} = .1 \Rightarrow x/5000 = \log_{.5} .1$ $\Rightarrow x = 5000 \log_{.5} .1 \approx 16610$ years.
4. Heating and cooling: $\overline{f(x)} =$	Q: How warm after __ minutes? Substitute:	Q: How long to reach __ degrees? Solve:
	Use an engineering calculator.	Use a spreadsheet.

The solutions of the inverse problems are simply applications of the definition of the logarithm!

Exercise 5.4.13

Complete the row for the heating/cooling model.

A different, but related, set of problems is about *finding the parameters* of an exponential model described indirectly. Specifically, there are four quantities in the equation:

$$y = Ca^x\,.$$

If *three* of them are known, find the fourth.

Example 5.4.14: compounded interest, original deposit = ?

Suppose I need \$10,000 in 5 years. How much do I need to deposit assuming that the APR will be 10%? Our model produces the following:

$$10,000 = C \cdot 1.1^5\,.$$

We solve for C:

$$C = \frac{10,000}{1.1^5} \approx \$6,209\,.$$

Example 5.4.15: compounded interest, past balance=?

Suppose I have \$1000. How much did I have 3 years ago if the APR has been 10%? Instead of thinking of the unknown to be the original deposit at time $x = 0$, we assume that today is $x = 0$ and we look backward in time, $x = -3$. Then our model produces the following:

$$1000 = C \cdot 1.1^{-3}\,.$$

We solve for C:

$$C = 1000 \cdot 1.1^{-3} \approx \$750\,.$$

Example 5.4.16: compounded interest, APR = ?

What APR will double my money in 5 years? If a is the rate of growth, our model produces the following:

$$2 = 1 \cdot a^5\,.$$

We solve for a:

$$a = \sqrt[5]{2} = 2^{1/5} \approx 1.15\,.$$

The APR then should be at least 15%.

Exercise 5.4.17

What has been my APR if I have tripled my money in 20 years?

Example 5.4.18: radiocarbon dating, half-life = ?

What is the half-life of a radioactive element if an experiment has shown that it loses .1% of its weight in 1 year? The model is as follows:

$$.999 = 1 \cdot a^1\,.$$

Therefore,

$$a = .999\,.$$

We have now a complete model:

$$y = C \cdot .999^x\,.$$

If x is the half-life, then:

$$.5 = 1 \cdot .999^x \implies \ln .5 = \ln(.999^x) = x \ln .999\,.$$

Solve for x:

$$x = \frac{\ln .5}{\ln .999} \approx 690 \text{ years.}$$

Exercise 5.4.19

What is the half-life of a radioactive element if an experiment has shown that it loses 1% of its weight in 10 years?

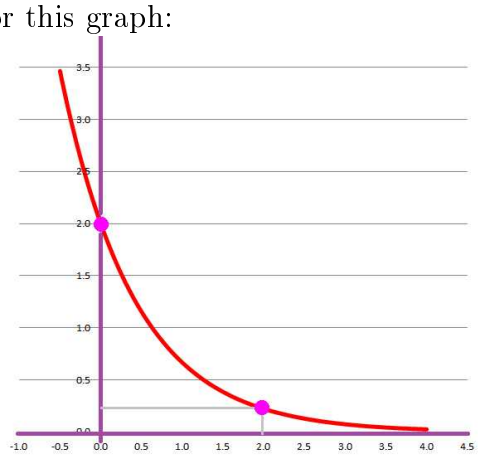
It is also possible that only *two out of four* quantities in the equation are known, but they are know twice:

$$y_1 = Ca_1^x \text{ and } y_2 = Ca_2^x\,.$$

Then the missing parameters can still be found.

Example 5.4.20: exponential function from graph

Find a representation $y = Ca^x$ for this graph:

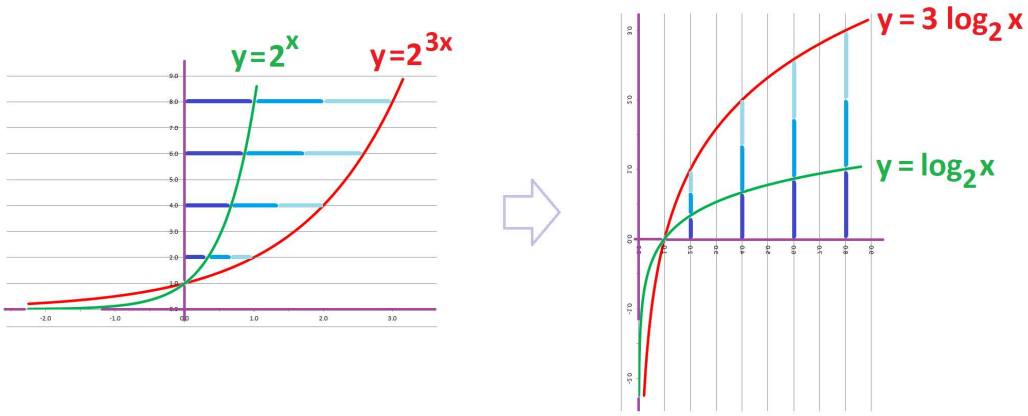


Pick two points on the graph and use them to write two equations for $f(x) = Ca^x$:

- The point $(0, 2)$ is on the graph. $\implies x = 0,\ y = 2 \implies Ca^0 = 2 \implies C = 2\,.$
- The point $(2, .25)$ is on the graph. $\implies x = 2,\ y = .25 \implies 2a^2 = .25 \implies a = \sqrt{.125}\,.$

Just like with a straight line, we only need two points!

We observed in the last chapter that we can get all exponential functions by horizontally stretching and flipping a single one of them:



It follows that we can get all logarithms by vertically stretching and flipping a single one of them.

Let's prove this fact with algebra. Suppose $a > 0$ and $b > 0$ are two bases that we want to match. Let's transform the former into the latter:

$$\begin{aligned} y &= a^x && \text{Use one of the Cancellation Laws to insert } b^{\log_b}. \\ &= b^{\log_b a^x} && \text{Now the Multiplication-Exponentiation Rule of Logarithms.} \\ &= b^{(\log_b a)\cdot x} && \text{But } \log_b a \text{ is just a constant coefficient!} \end{aligned}$$

We conclude the following:

- The graph of $y = a^x$ is that of $y = b^x$ stretched *horizontally* by a factor of $\log_b a$.
- The graph of $y = \log_a x$ is that of $y = \log_b x$ shrunk *vertically* by a factor of $\log_b a$.

The algebraic interpretation is below.

Theorem 5.4.21: Exponent and Logarithm Base Conversion Formulas

For every pair $a, b > 0$ and every $x > 0$, we have:

$$\begin{aligned} a^x &= b^{(\log_b a)\cdot x} \\ \log_a x &= \frac{1}{\log_b a} \log_b x \end{aligned}$$

Exercise 5.4.22

Derive the second formula.

The following is an important conclusion:

- Every exponential function can be *easily* expressed in terms of any other.
- Every logarithm function can be *easily* expressed in terms of any other.

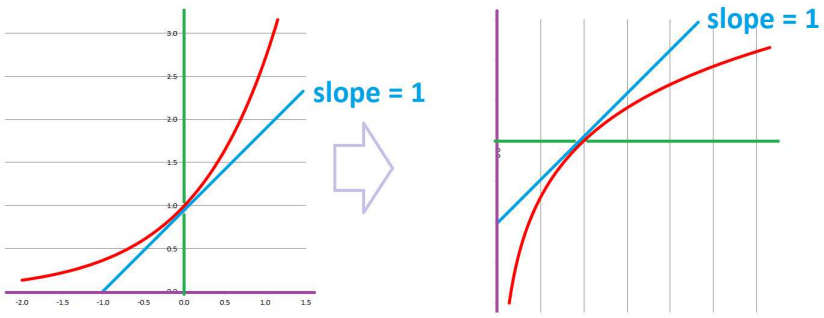
So, just as we need only one quadratic polynomial, $y = x^2$, we only need *one* exponential function and one logarithm!

Definition 5.4.23: natural logarithm

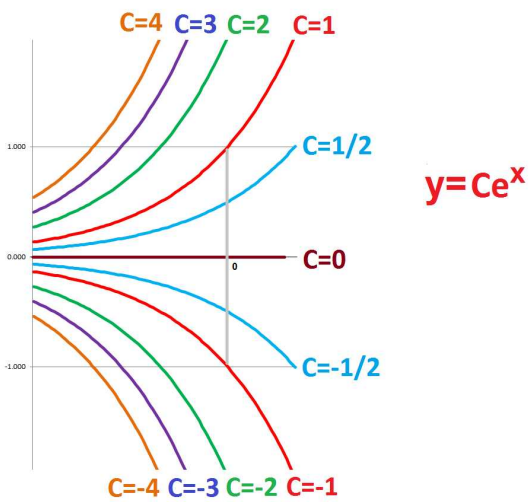
The logarithm base e is called the *natural logarithm* and denoted by

$$y = \ln x$$

Its graph crosses the x-axis at 45 degrees (to be discussed in Chapter 2DC-3):



What we have discovered is that it suffices to consider a single base – the *standard* base – to get *all* of the exponential models! Indeed, by making *multiples of* $y = e^x$, we can get all possible patterns of exponential growth:



Exercise 5.4.24

Show that the vertical stretches of the exponential functions can be expressed as horizontal shrinks. Same for the logarithms.

Definition 5.4.25: exponential model

The *exponential model* (in the standard form) is a function:

$$y = Ce^{kx}$$

with k called its *rate*. Furthermore,

- $y = Ce^{kx}$, $k > 0$, is called the *exponential growth model*, and
- $y = Ce^{kx}$, $k < 0$, is called the *exponential decay model*.

The transition

$$\text{from } \; y = Ca^x \; \text{ to } \; y = Ce^{kx}\,,$$

is straight-forward:

$$Ca^x = Ce^{kx} \implies a^x = e^{kx} = \left(e^k\right)^x \implies a = e^k\,.$$

Example 5.4.26: population growth

Suppose that the population grows by 10% a year. How long will it take to double?

In contrast to the previous analysis of the same problem, we will find the *standard*, base e, exponential

model for this setup (x time):

$$f(x) = Ce^{kx}, \quad \text{what is } k?$$

Let's assume $f(0) = 1$, then $f(1) = 1.1$. It follows that $C = 1$. Substitute these into both models:

$y = a^x$	$y = e^{kx}$	
	$e^{k\cdot 1} = 1.1$	$\Longrightarrow$
$a = 1.1$	$k = \ln 1.1\,.$	

We can use either of the models to solve the original problem:

$y = 1.1^x$	$y = e^{\ln 1.1\cdot x}$	
$1.1^x = 2$	$e^{\ln 1.1\cdot x} = 2$	$\Longrightarrow$
$x = \log_{1.1} 2$	$x = \dfrac{\ln 2}{\ln 1.1} \approx 7.27\,.$	

Example 5.4.27: find inverse from formula

Find the inverse of

$$f(x) = e^{x^3}\,.$$

Solve the equation: $y = e^{x^3}$, solve for x. To get to x, we need to get rid of the exponent. To cancel it, apply its inverse to both sides. For $y = e^x$, it's $x = \ln y$:

$$\begin{aligned} \ln y &= \ln e^{x^3} \\ \ln y &= x^3 \end{aligned}$$

Apply the inverse to get to x:

$$\sqrt[3]{\ln y} = \sqrt[3]{x^3} = x\,.$$

Answer: $f^{-1}(y) = \sqrt[3]{\ln y}$.

In summary:

$$y = Ca^x \;\to\; \boxed{\begin{array}{l} a = e^k \\ k = \ln a \end{array}} \;\to\; y = Ce^{kx}$$

Below are the matched up lists of the formulas most frequently used:

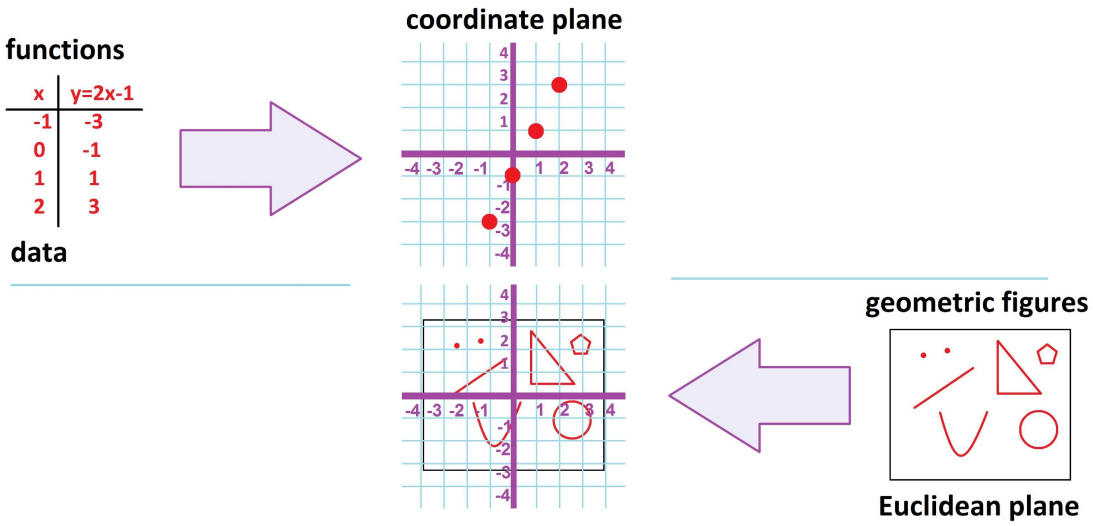
Theorem 5.4.28: Natural Log and Exp Formulas

The following formulas hold for every pair x and y for which they are defined:

$e^{\ln x} = x$	$\ln e^x = x$
$e^0 = 1$	$\ln 1 = 0$
$e^x e^y = e^{x+y}$	$\ln x + \ln y = \ln xy$
$\dfrac{e^x}{e^y} = e^{x-y}$	$\ln x - \ln y = \ln \dfrac{x}{y}$
$\left(e^x\right)^y = e^{xy}$	$\ln x^y = y \ln x$

5.5. The Cartesian system for the Euclidean plane

In Chapter 2, we introduced the Cartesian coordinate plane as a device for visualizing *functions*; it's where the graphs live (top row):



We now approach it from another direction (bottom row): We want to study the *Euclidean plane*, and Euclidean geometry, algebraically. We *superimpose* – as if it is drawn on a transparent piece of plastic – the Cartesian grid over this plane.

The main difference is that the former plane doesn't need to have a square grid as the units of x and y might be unrelated (dollars vs. hours). The latter does and the units of x and y are better be those of length (miles, feet, etc.).

The idea of "analytic geometry" is to use a coordinate system to transition

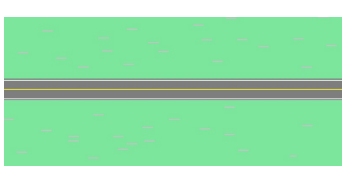
- from *geometry*: points, then lines, triangles, circles, then planes, cubes, spheres, etc.,
- to *algebra*: numbers, then combinations of numbers, then relations and functions, etc.

This will allow us to solve geometric problems *without measuring* – because everything is pre-measured! We will initially limit ourselves to the two simplest geometric tasks:

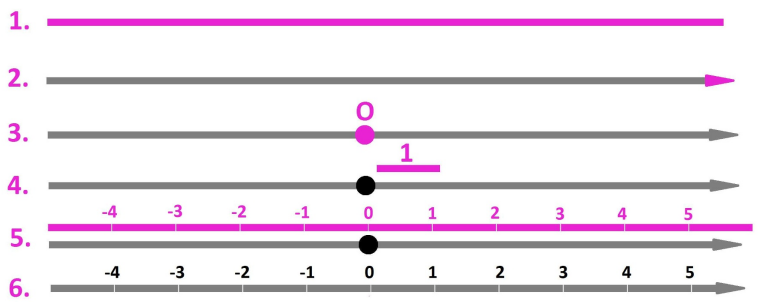
1. finding distances: between two points, between a point and a line or curve, etc., and
2. finding angles: between two lines or two curves.

We start with *dimension* 1. Let's repeat the construction from Chapter 1 first.

Suppose we live on a *road* surrounded by nothingness:



The coordinate system intended to capture what happens on this road is devised to be superimposed on the road.



It is built in several stages:

1. Draw a line, the *x-axis*.
2. Choose one of the two directions on the line as *positive*, then the other is *negative*.
3. Choose a point O as the *origin*.
4. Set a segment of the line – of length 1 – as a *unit*.
5. Use the segment to measure distances to locations from the origin O – positive in the positive direction and negative in the negative direction – and add marks to the line, the *coordinates*; later the segments are further subdivided to fractions of the unit, etc.
6. We have a coordinate system on the line.

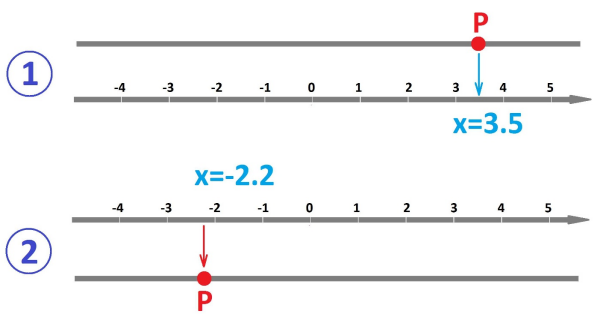
The result is a correspondence:

location P ⟷ number x

This is the real line.

The correspondence works in both directions.

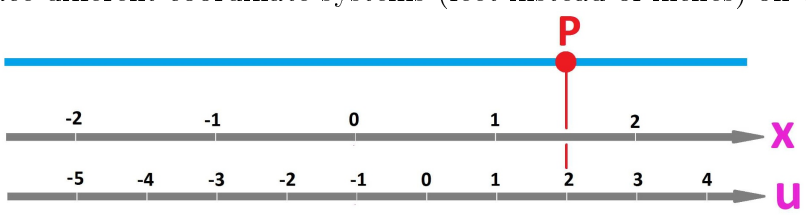
For example, suppose P is a *location* on the line. We then find the distance from the origin – positive in the positive direction and negative in the negative direction – and the result is the coordinate of P, some *number* x. We use the nearest mark to simplify the task.



Conversely, suppose x is a *number*. We then measure x as the distance to the origin – positive in the positive direction and negative in the negative direction – and the result is a *location* P on the line. We use the nearest mark to simplify the task.

Example 5.5.1: different coordinate systems, dimension 1

Of course, we can place different coordinate systems (feet instead of inches) on the same line:



Here, the point P shown has:

- coordinate 1.5 according to the first system and

- coordinate 2 according to the other.

We know from Chapter 3 that all the functions defined on the first axis are transformed to the ones on the second by a single function; in this particular case, it is:

$$x = (u + 1)/2\,.$$

Of course, all the functions defined on the second axis are transformed to the ones on the first by the *inverse* of this function:

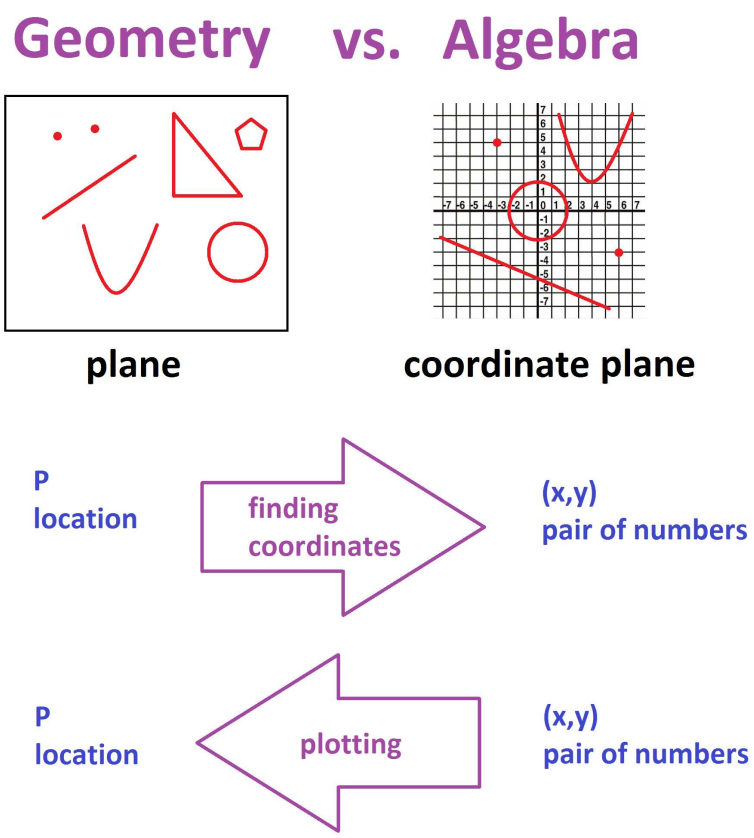
$$u = 2x + 1\,.$$

With these two functions, all the quantities – geometric or physical – defined within the two coordinate systems are transformed to each other.

Exercise 5.5.2

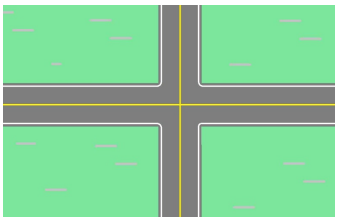
What transformation isn't mentioned above? Provide an illustration and a formula for the combination of the three.

Now the coordinate system for *dimension* 2, the plane. There is much more going on than before:

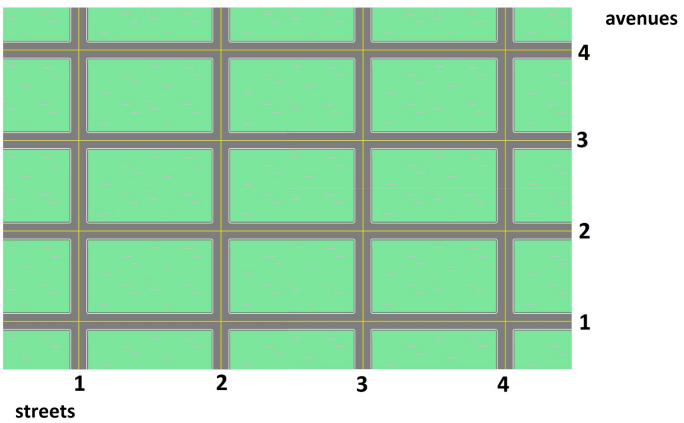


The idea is the same: solving geometric problems with algebra.

Let's repeat – with some minor changes – the construction from Chapter 2 first. Suppose we live on a *field* and we build two roads intersecting at 90 degrees:

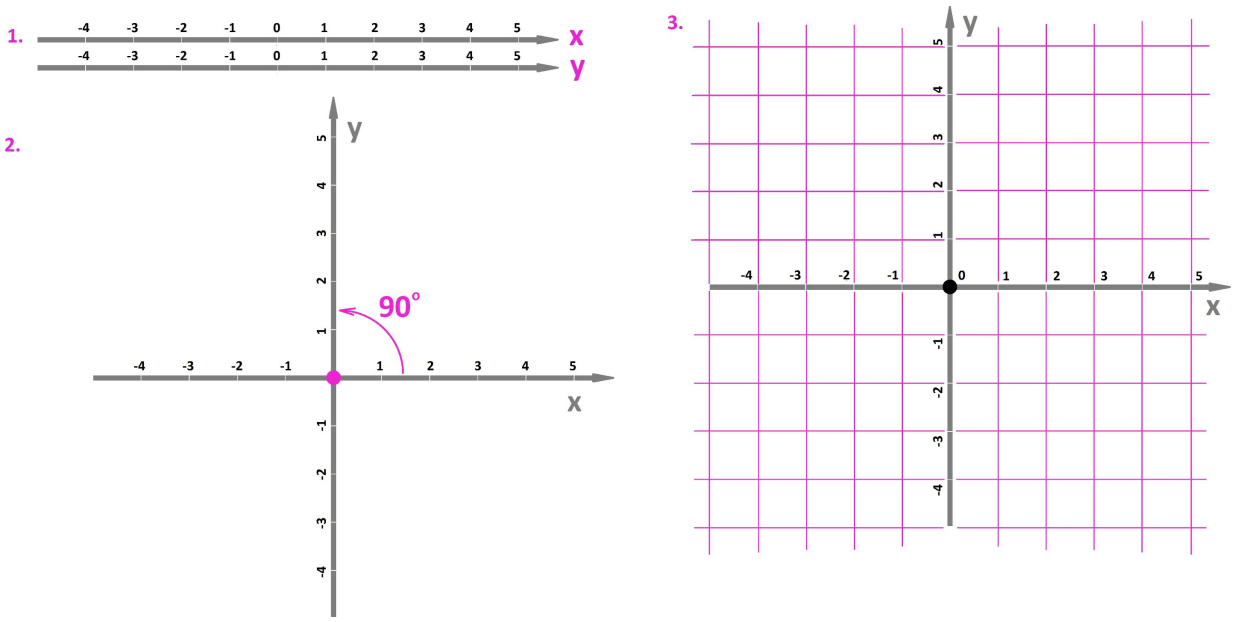


We can then treat either of the two roads as a 1-dimensional Cartesian system, as above, and use their milestones to navigate. But what about the rest of the field? How do we navigate it? We could build a city with a grid of streets:



We also number the streets. It's two-dimensional now!

A new coordinate system intended to capture what happens in this city or on this field is devised to be superimposed on the field. It's a grid:



It is built in several stages:

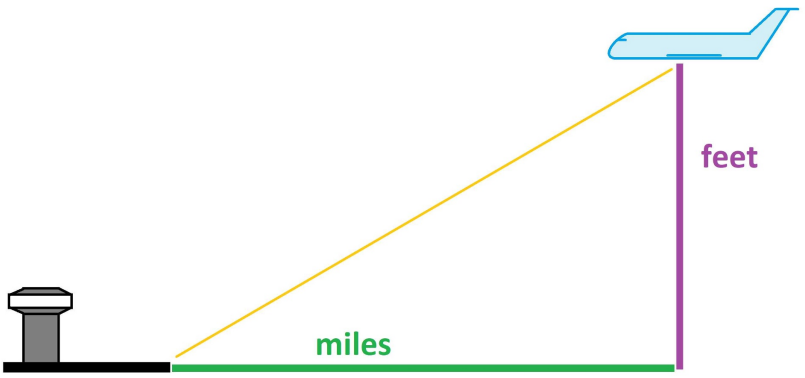
1. Choose two identical coordinate axes, the *x-axis* first and the y-axis second, with the same units.
2. Put the two axes together at their origins so that it is a 90-degree turn from the positive direction of the x-axis to the positive direction of the y-axis.
3. Use the marks on the axes to draw a grid.

Warning!

The xy-plane isn't the same as the yx-plane.

Example 5.5.3: units

It is possible though uncommon to have different units for the two axes:



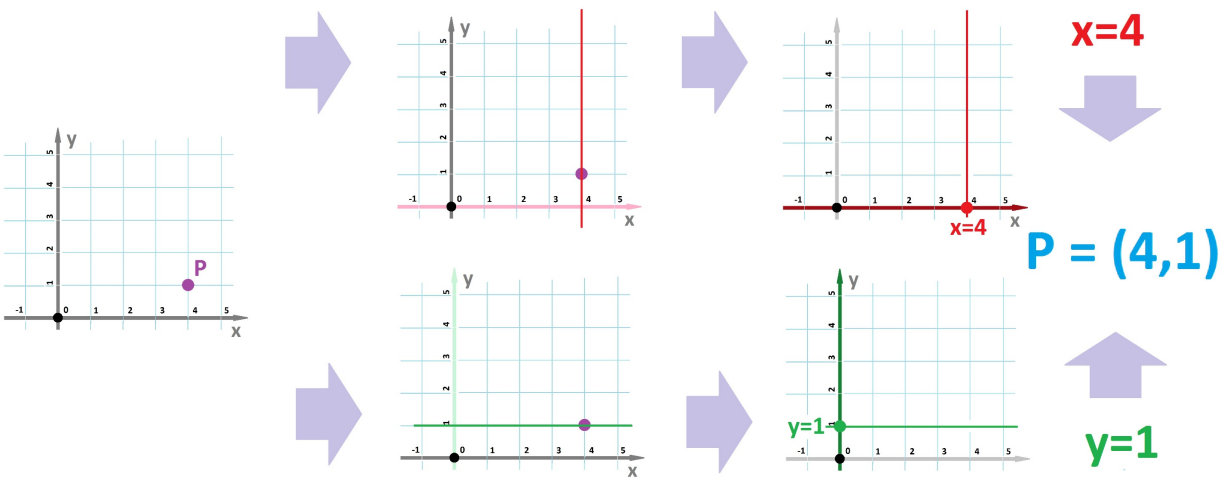
We have a correspondence that works in *both directions*:

location **P** $\longleftrightarrow$ a pair of numbers **(x,y)**

Example 5.5.4: coordinates from point and back

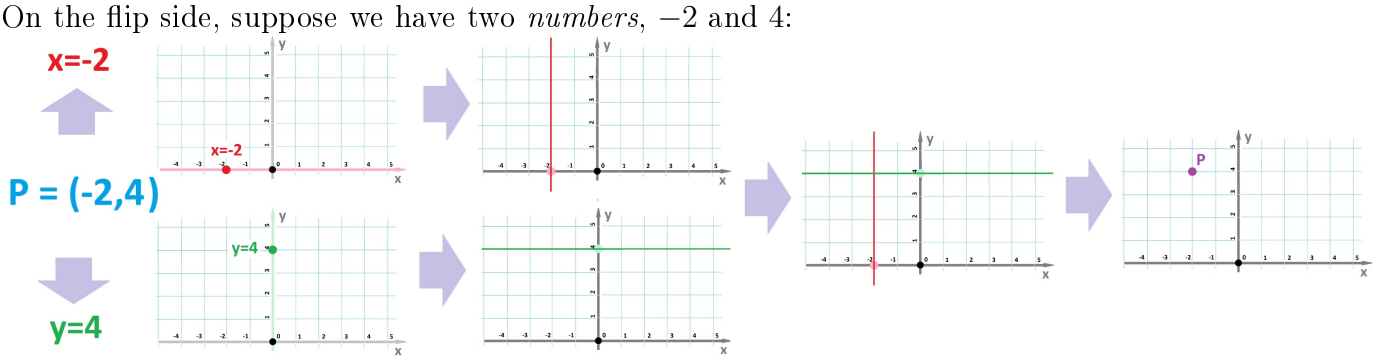
Suppose we have the Euclidean plane equipped with a Cartesian system.

Suppose P is a *point* as shown:



1. We draw a vertical line through P until it intersects the x-axis. The point of intersection then lies on this axis, which is equipped with a 1-dimensional Cartesian system. This point has a coordinate, say 4, within this system.
2. We draw a horizontal line through P until it intersects the y-axis. The point of intersection then lies on this axis, which is equipped with a 1-dimensional Cartesian system. The point has a coordinate, say 1, within this system.
3. We have discovered that our point has coordinates $P = (4, 1)$!

On the flip side, suppose we have two *numbers*, -2 and 4:



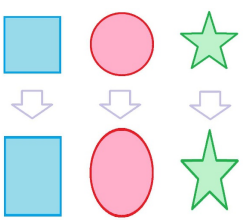
1. We find the location on the x-axis with coordinate -2. We then draw a vertical line through this point.

2. We find the location on the y-axis with coordinate 4. We then draw a horizontal line through this point.
3. The intersection of these two lines is the corresponding *point* $P = (4, 1)$ on the plane!

In summary:

- If P is a *location* on the plane, we find the distances from either of the two axes to that location – positive in the positive direction and negative in the negative direction – and the result is the two coordinates of P, some *numbers* x and y.
- If x and y are *numbers*, we measure x as the distance from the y-axis and y as the distance from the x-axis – positive in the positive direction and negative in the negative direction – and such locations together form a vertical line and a horizontal line, and the intersection of these two is the *location* P on the plane.

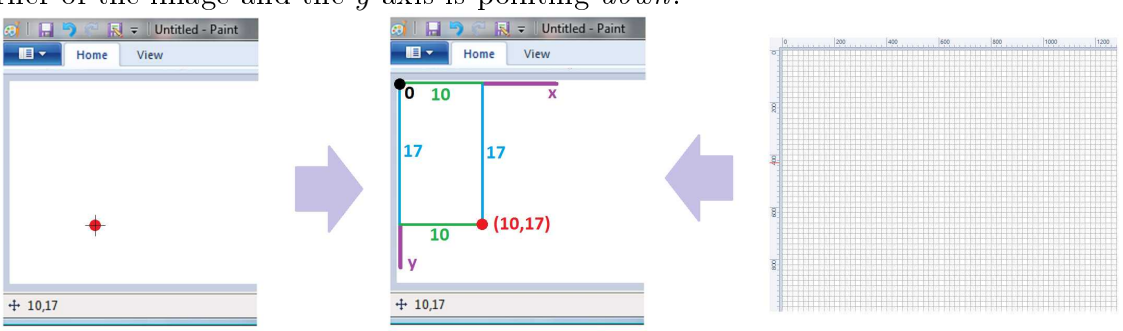
The main difference from the xy-plane we have seen previously comes from its purpose. In order to do *Euclidean geometry* on this coordinate plane, squares should be squares and not rectangles; circles should be circles and not ovals, etc.:



Even though we can make larger or smaller plots, the relative dimensions (and, consequently, the *angles*) will have to remain the same. To avoid such a disproportional resizing, we need to make sure that we use the *same units* for the x-axis and the y-axis. This will have to be a *square grid*. Thus, we are just narrowing down the scope of possibilities in comparison to the way the Cartesian system was treated previously.

Example 5.5.5: coordinates used in computing

The 2-dimensional Cartesian system isn't as widespread as the one for dimension 1 (numbers). It is, however, common in certain areas of computing. For example, drawing applications allow you to make use of this system – if you understand it. The location of your mouse is shown in the status bar on the lower left, constantly updated in real time. The main difference is that the origin is in the left upper corner of the image and the y-axis is pointing *down*:

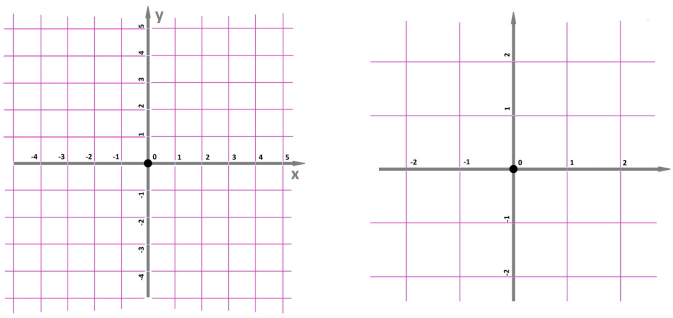


The choice is explained by the way we write: downward.

Example 5.5.6: different coordinate systems, dimension 2

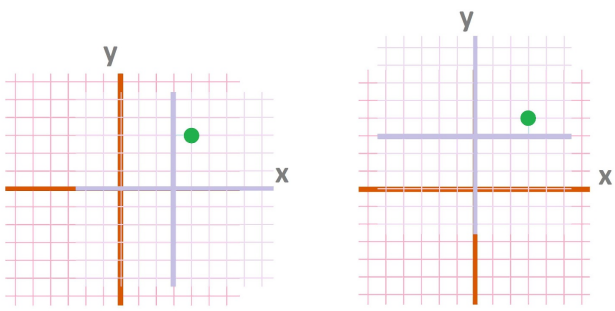
Of course, we can place different coordinate systems on the same plane.

For example, we can have two systems that differ only by scale:



That's a uniform stretch!

Or they can differ only by the location of the origin:



That's a shift!

We saw in Chapter 3 how we can transfer information (points, set, functions, etc.) defined on the first plane to the second. It only takes a single function. For the former example, it is

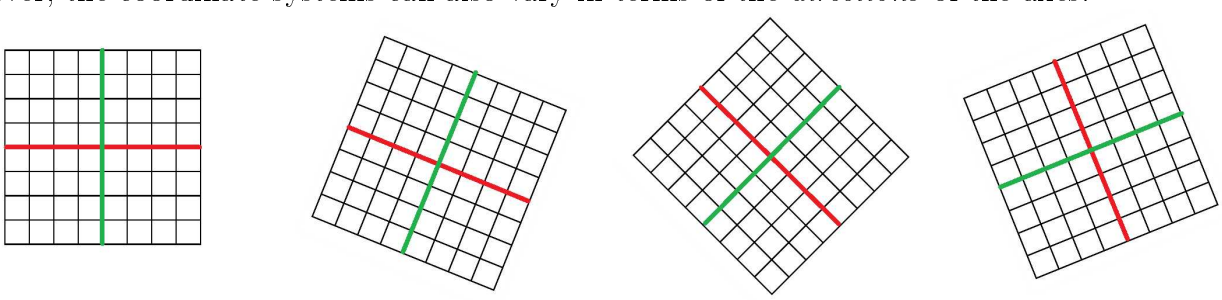
$$(x, y) \mapsto (x/2, y/2)\,.$$

For the latter example, it is

$$(x, y) \mapsto (x - 3, y - 4)\,.$$

With these functions, all the quantities – geometric or physical – defined within the two coordinate systems are transformed to each other.

Moreover, the coordinate systems can also vary in terms of the *directions* of the axes:



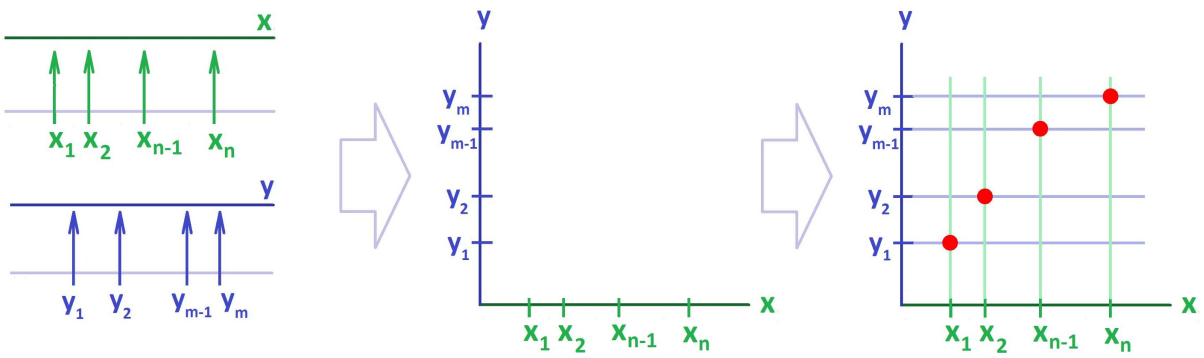
Rotations and other transformations of the plane are considered in Chapter 5DE-2.

Exercise 5.5.7

What transformation isn't mentioned above? Provide an illustration and a formula for its combination with the first two.

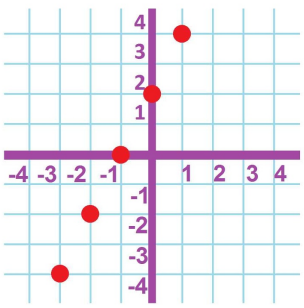
Example 5.5.8: motion on plane, parametric curves

A sequence of x's, x_n, and a sequence of y's, y_n, will create a *sequence of points on the plane*:



In contrast to the study in Chapter 1, we can represent the path of a moving ball beyond just "up or down"!

To begin with, when both sequences are arithmetic progressions, the path is a straight line:

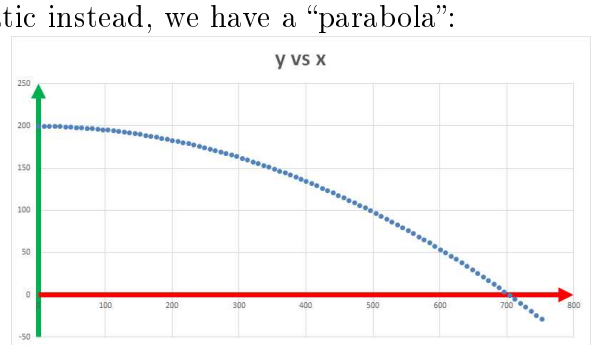


With n as *time*, we have, recursively:

$$x_{n+1} = x_n + p,\ y_{n+1} = y_n + q\,,$$

where the increments of the arithmetic progressions, p and q, represent the horizontal and the vertical components of the *velocity*, respectively (above $q = 1$, $p = 2$ with the initial values -1 and 0 as an example).

If we choose y_n to be quadratic instead, we have a "parabola":



We can now represent the *path of a thrown ball*. For example, from the elevation of 200 feet, we shoot a cannonball straight forward at 200 feet per second:

$$x_n = 200n,\ y_n = 200 - 16n^2\,.$$

Circular motion is considered later in this chapter.

Exercise 5.5.9

Show that the picture describes the path given by the sequences.

5.6. The Euclidean plane: distances

Since everything in the Cartesian system is pre-measured, we can solve some geometric problems by algebraically manipulating the coordinates of points.

In this section, we consider the very basic geometric task of computing – as opposed to measuring – *distances*.

First, the *line*.

The distance – a number – between two locations P and Q is inherited from the Euclidean line that underlies the Cartesian system. The distance between two points P and Q is denoted as follows:

Distance on line

$$d(P,Q)$$

Now, how do we express this number in terms of their coordinates, say x and x'?

Example 5.6.1: distance dim 1

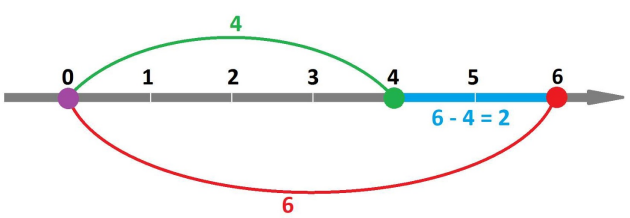
One finds the distance that has been covered on the road by *subtracting* the number on the milestone in the beginning and the number on the milestone at the end:

$$\text{from } P=4 \text{ to } Q=6 \implies \text{distance } = Q-P = 6-4 = 2\,.$$

But what if we are moving in the opposite direction? The distance should be the same! And the computation should be the same:

$$\text{from } Q=6 \text{ to } P=4 \implies \text{distance } = Q-P = 6-4 = 2\,.$$

Indeed:



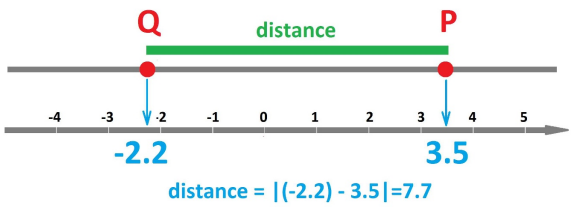
In other words, one must subtract the smaller number from the larger one every time in order for the computation to make sense.

Theorem 5.6.2: Distances on Line

The distance between two locations P and Q on the real line given by their coordinates x and x' is:

- $d(P,Q) = x' - x$ *when* $x < x'$,
- $d(P,Q) = x - x'$ *when* $x > x'$,
- $d(P,Q) = 0$ *when* $x = x'$.

Is there a *single formula* for this computation? The idea that the distance between two locations can never be negative suggests that this has something to do with the absolute value:



The *absolute value* function is defined to be

$$|a| = \begin{cases} -a & \text{when } a < 0\,, \\ 0 & \text{when } a = 0\,, \\ a & \text{when } a > 0\,. \end{cases}$$

We just substitute $a = x' - x$ into this formula to prove the following:

Theorem 5.6.3: Distance Formula for Dimension 1

The distance between two points on the real line with coordinates x and x' is the absolute values of their difference (in either order).

In other words, we have:

$$d(x, x') = |x - x'| = |x' - x| = d(x', x)$$

Exercise 5.6.4

Derive the formula from the following:

$$d(P, Q) = \begin{cases} x - x' & \text{when } x > x'\,, \\ 0 & \text{when } x = x'\,, \\ x - x' & \text{when } x < x'\,. \end{cases}$$

Exercise 5.6.5

Prove that for any two numbers x, x', we have $|x + x'| \leq |x| + |x'|$.

Exercise 5.6.6

Prove that the point half-way between points $P = x$ and $Q = x'$ (called their "midpoint") has the coordinate $\dfrac{x + x'}{2}$.

The word "stretch" that we have used in the past now takes a meaning that relies on the idea of distance: The distances increase proportionally.

Theorem 5.6.7: Linear Transformations in Dimension 1

A linear function stretches the x-axis by a factor of $|m|$, where m is its slope.

Proof.

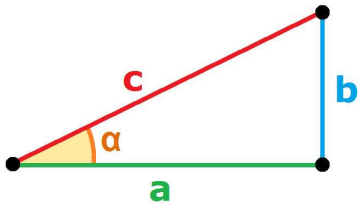
This is what happens to the distance between two points u and v, which is $|v - u|$, after a linear function $f(x) = mx + b$ is applied:

$$|f(v) - f(u)| = |(mv + b) - (mu + b)| = |mv - mu| = |m| \cdot |v - u|\,.$$

The distance has increased by a factor of $|m|$. (We say that it has decreased by a factor of m when $|m| < 1$). This stretch/shrink factor is the same everywhere.

Next, the *plane*.

We start with a result that is one of the most important.



Theorem 5.6.8: Pythagorean Theorem

Suppose we have a right triangle with sides a, b, c, with c the longest one facing the right angle. Then, we have the following:

$$a^2 + b^2 = c^2$$

Proof.

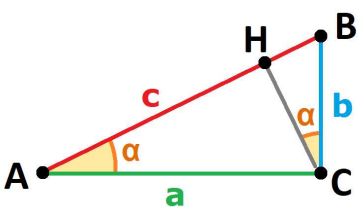
We use what we know about similar triangles (i.e., the ones with equal angles):

▶ The ratio of any two corresponding sides of similar triangles is the same.

Let ABC be our right triangle, with:

- vertex A opposite to side a,
- vertex B opposite to side b,
- vertex C opposite to side c.

We draw the height (the line perpendicular to c) from C, and call H its intersection with the side c:



The new triangle ACH is similar to our original triangle ABC, because they both have a right angle, and they share the angle at A, α. In the same way, we prove that the triangle CBH is also similar to ABC. The similarity of these two pairs of triangles leads to the equality of ratios of the corresponding sides:

$$\begin{array}{l|l} \dfrac{BC}{AB} = \dfrac{BH}{BC} & \dfrac{AC}{AB} = \dfrac{AH}{AC} \\ BC^2 = AB \cdot BH & AC^2 = AB \cdot AH \end{array}$$

We add the two items in the last row and factor:

$$b^2 + a^2 = BC^2 + AC^2 = AB \cdot BH + AB \cdot AH = AB \cdot (AH + BH) = AB^2 = c^2 \,.$$

Exercise 5.6.9

State the converse of the theorem. Is it true?

Exercise 5.6.10

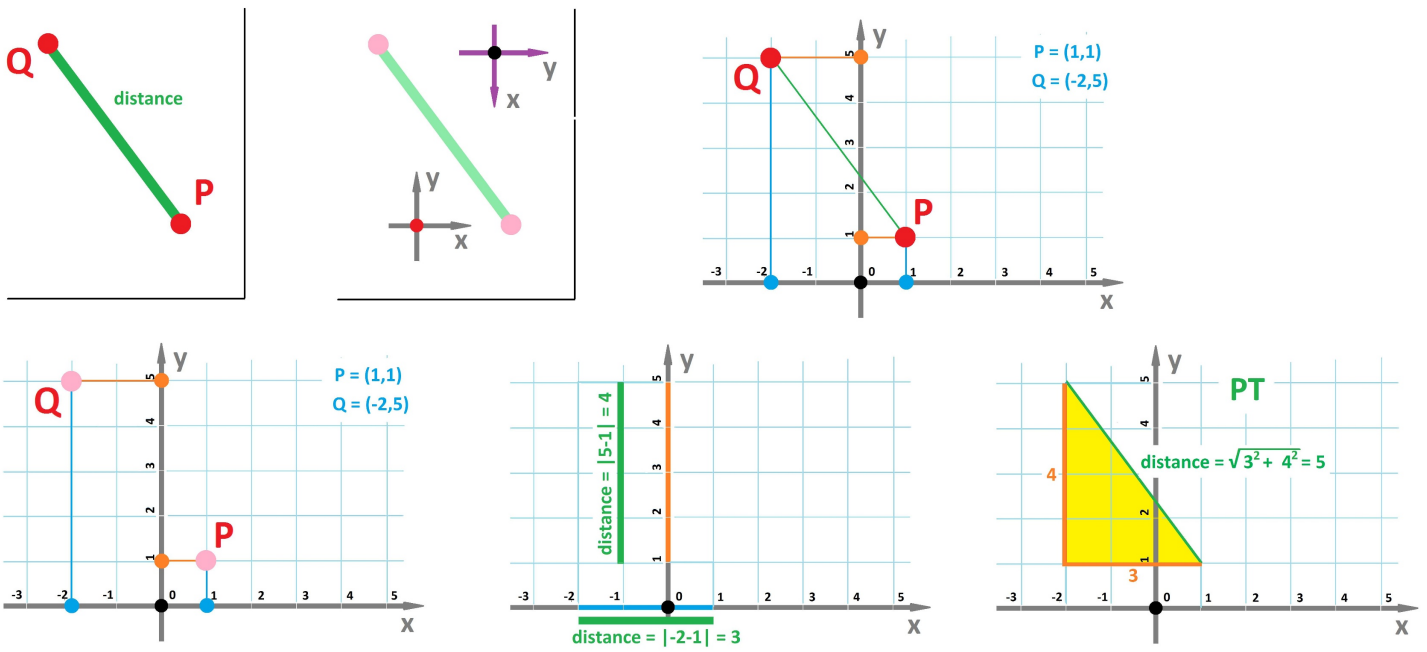
What do the ratios in the proof tell us about the trigonometric functions of the angle α?

The value of the distance between two locations P and Q is inherited from the Euclidean plane that underlies the Cartesian system. The distance between two points P and Q is denoted as follows:

Distance on plane

$d(P, Q)$

Now, what if this time we have a Cartesian system placed on top of this piece of paper? How do we express the distance between P and Q in terms of their coordinates (x, y) and (x', y')? The idea is to find the distances along the axes first:



We applied the *Distance Formula for Dimension* 1 for either of the two axes and then used the *Pythagorean Theorem*.

The following is one of the most useful results in geometry of the Cartesian plane.

Theorem 5.6.11: Distance Formula for Dimension 2

The distance between two points with coordinates $P = (x, y)$ and $Q = (x', y')$ is

$$d(P, Q) = \sqrt{(x - x')^2 + (y - y')^2}$$

Proof.

According to the formula:

- The distance between x and x' on the x-axis is $|x - x'|$.
- The distance between y and y' on the y-axis is $|y - y'|$.

Then, the segment between the points $P(x, y)$ and $Q = (x', y')$ is the hypotenuse of the right triangle with sides: $|x - x'|$ and $|y - y'|$. Then our conclusion below follows from the *Pythagorean Theorem*:

$$d(P, Q)^2 = |x - x'|^2 + |y - y'|^2 .$$

Since $|z|^2 = z^2$ for any z, we can remove the absolute value signs.

The result is so important that one can even say that the 90-degree angle between the axes was chosen so that we can produce this formula from the Pythagorean theorem.

Exercise 5.6.12

Find the distance between the points $(-5, 2)$ and $(2, -1)$.

Warning!

Combine x's with x's and y's with y's.

This is how the two formulas for the two dimensions can be matched up:

	distances				
dimension			x-axis		y-axis
1	$d(P,Q)^2$	$=$	$(x-x')^2$		
2	$d(P,Q)^2$	$=$	$(x-x')^2$	$+$	$(y-y')^2$

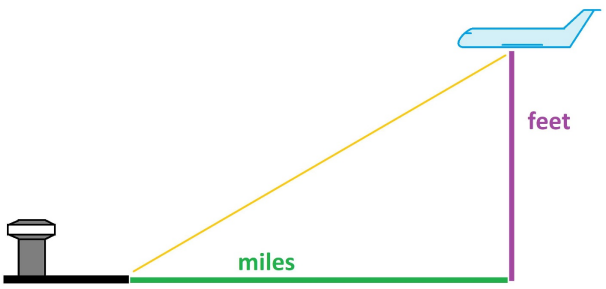
In other words:

- The square of the distance is the sum of the squares of the differences of the coordinates.

We will see a continuation of this list and of this pattern in higher dimensions (Chapter 3IC-4).

Example 5.6.13: miles and kilometers

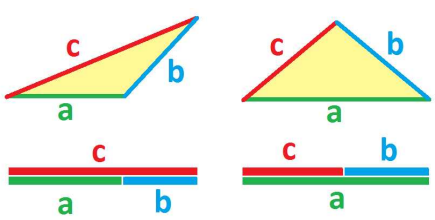
It is possible to have the x-axis measured in different units from the y-axis. For example, it is typical to measure the distance to the airport in miles but the altitude in feet:



Also, one can speak, hypothetically, of a point located "2 miles east and 5 kilometers north" from here. However, when this is the case, the Distance Formula won't be applicable anymore!

Here is a familiar property of the lengths of the sides of a triangle: The length of any side is less than the sum of the lengths of the other two. In other words, if a, b, c are these three sides, then:

$$c < a + b\,.$$



As you can see (bottom row), even when the triangle "degenerates" to a segment, we have an equation:

$$c = a + b\,.$$

Thus, the inequality remains true though non-strict. We restate this extended inequality as follows.

Theorem 5.6.14: Triangle Inequality

For any three points P, Q, R on the plane, we have the following:

$$d(P,R) \leq d(P,Q) + d(Q,R)$$

In other words, the straight line is the shortest.

Exercise 5.6.15

Derive the theorem from the Distance Formula.

Exercise 5.6.16

Derive the inequality for a right triangle from the Pythagorean Theorem.

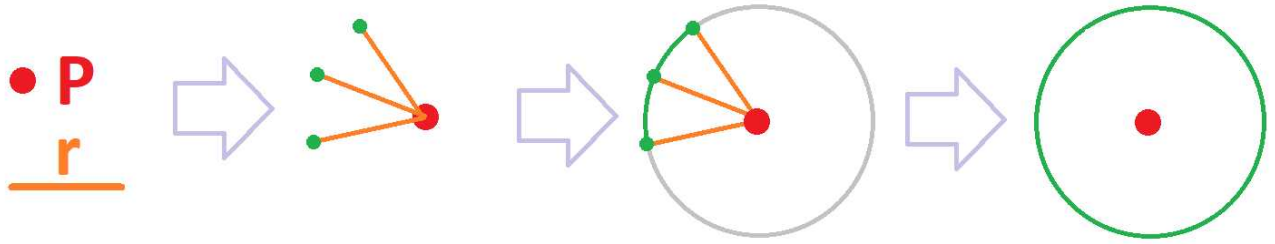
When the triangle degenerates into a segment, we have the triangle inequality for dimension 1 presented above.

Exercise 5.6.17

Prove that the point M half-way between points $P = (x, y)$ and $Q = (x', y')$ – called their *midpoint* – is given by their average coordinates:

$$M = \left(\frac{x + x'}{2}, \frac{y + y'}{2} \right).$$

The simplest geometric figure that relies on the idea of distance is the *circle*. If the center, which is a point on the plane, and the radius, which is a positive number, are given, we just draw segments of this length in all directions from the center and their ends will create the circle:



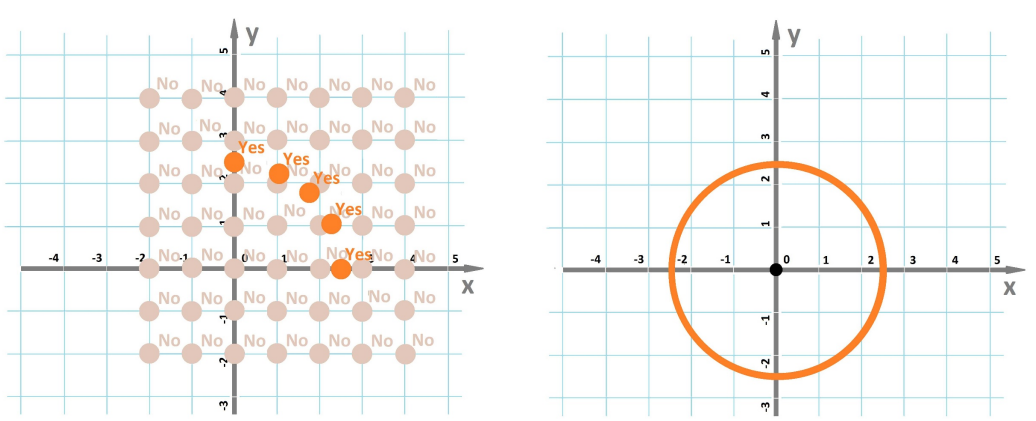
A more precise way to represent a circle is as a subset of the plane defined via the set-building notation, as follows.

Definition 5.6.18: circle

The *circle* of radius $r > 0$ centered at point P on the plane is the set of all points r units away from P:

$$\{Q : \ d(P, Q) = r\}.$$

We check one point at a time:



Warning!

As a set, the circle does not contain the center of the circle nor the rest of its interior. It's a curve!

Exercise 5.6.19

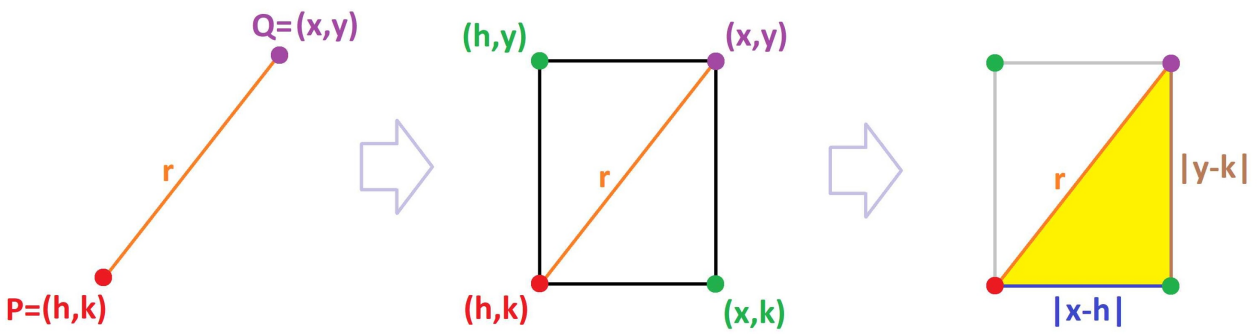
Prove that if the distance between the centers of two circles is larger than the sum of their radii, then the circles don't intersect. Hint: The Triangle Inequality.

What about a *coordinate representation* of circles?

Suppose $P = (h, k)$ is the center. Then, if a point $Q = (x, y)$ lies on the circle of radius r centered at P, then, according to the *Pythagorean Theorem*, or the *Distance Formula*, its distance to the center P is the following:

$$(x - h)^2 + (y - k)^2 = d(P, Q)^2 = r^2 .$$

The three quantities are visible below as we check point Q:



Conversely, if the point is not on the circle, the equation fails.

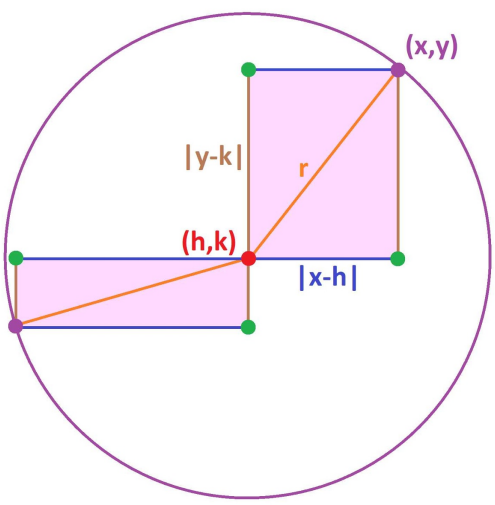
We arrive at the following conclusion.

Theorem 5.6.20: Equation of Circle

The circle of radius $r > 0$ centered at a point $P = (h, k)$ is the graph of the following relation:

$$(x - h)^2 + (y - k)^2 = r^2$$

In other words, every pair of numbers x and y that satisfy this equation produces a point (x, y) that lies on this circle, and vice versa. The circle is constructed from all possible rectangles like this:



So, the circle of radius $r > 0$ centered at a point $P = (h, k)$ is the following set:

$$\{(x, y) : \ (x - h)^2 + (y - k)^2 = r^2 \ \}.$$

The definition of circle can be violated in two ways: the distance to the center is too large or too small.

Definition 5.6.21: open and closed disks

- The *open disk* of radius $r > 0$ centered at point P on the plane is the set of all points less than r units away from P:

$$\{Q : \ d(P, Q) < r\}.$$

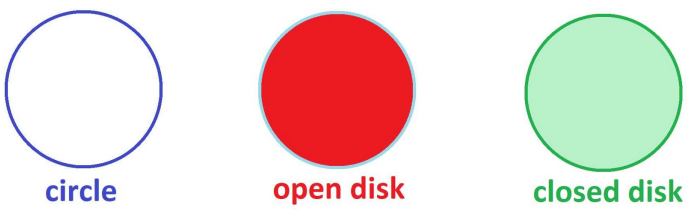
- The *closed disk* of radius $r > 0$ centered at point P on the plane is the set of all points less than or equal to r units away from P:

$$\{Q : \ d(P, Q) \leq r\}.$$

We also conclude the following:

- The open disk is the set of points that lie inside the circle of radius $r > 0$ centered at point P.
- The closed disk is set of all points that lie inside the circle of radius $r > 0$ centered at point P together with the circle itself.

So, the closed disk is the union of the open disk and the circle itself:



Now their coordinate representations.

Corollary 5.6.22: Disk via Inequality

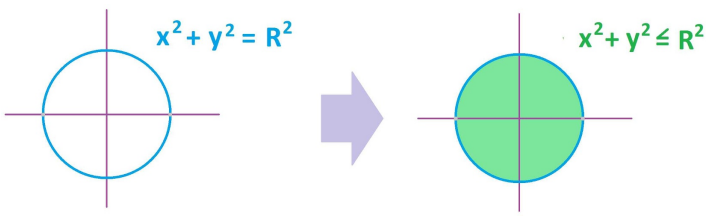
A point (x, y) belongs to the open disk of radius $r > 0$ centered at point $P = (h, k)$ if and only if it satisfies the inequality:

$$(x - h)^2 + (y - k)^2 < r^2$$

A point (x, y) belongs to the closed disk of radius $r > 0$ centered at point $P = (h, k)$ if and only if it satisfies the inequality:

$$(x - h)^2 + (y - k)^2 \leq r^2$$

This is the connection between the circle and the corresponding closed disk:



Exercise 5.6.23

Prove the theorem.

Exercise 5.6.24

Describe the relations present in the theorem.

Exercise 5.6.25

Write the three using the set-building notation.

So, it's just a matter of the distance from the center equal to r or less than (or equal) to r.

Example 5.6.26: circle via center and radius

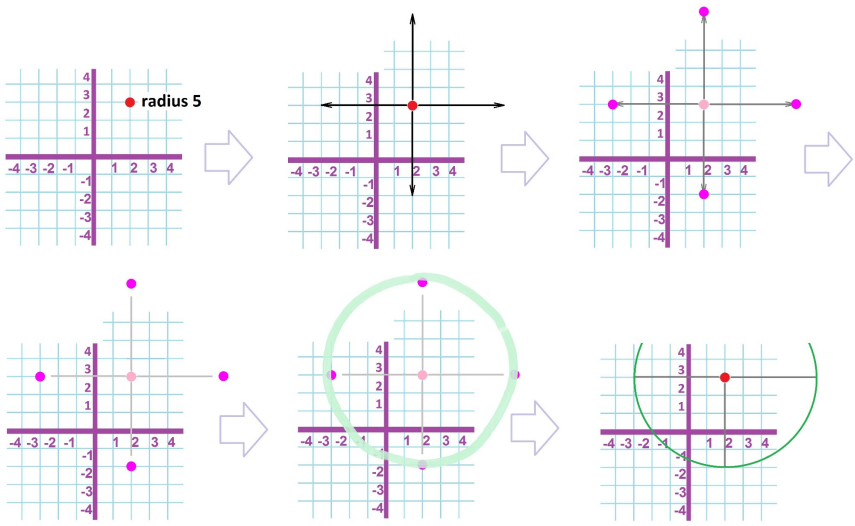
What is the equation of the circle centered at $(2, 3)$ and radius 5? This is just a matter of substitution:

$$h = 2,\ k = 3,\ r = 5\,.$$

Then, the theorem produces the following equation:

$$(x-2)^2 + (y-3)^2 = 5^2\,.$$

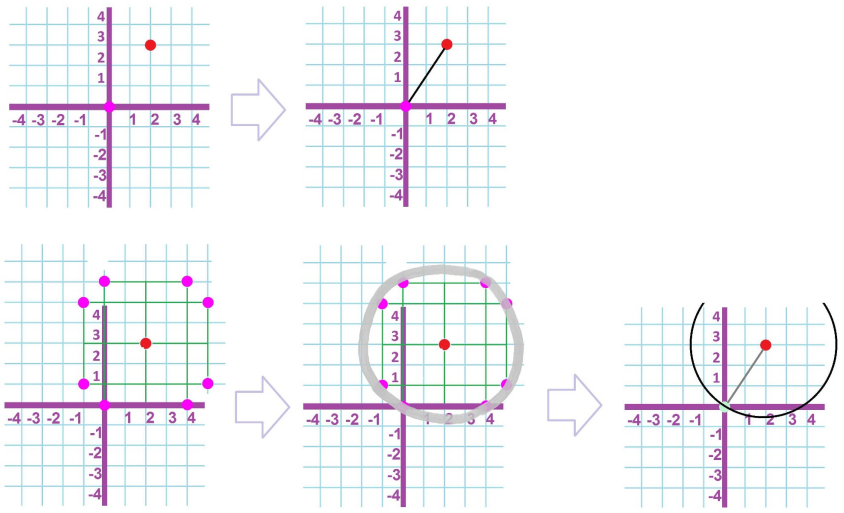
We plot to confirm:



In order to sketch a circle by hand, we can just make a step r units long in each of the four directions and then connect these points by a curve.

Example 5.6.27: circle via center and point

Sometimes the description of the circle is less explicit: What is the equation of the circle centered at $(2, 3)$ and passing through the origin? The center is known but the radius is missing! To find it, we observe that the distance from any point on the circle to its center is equal to its radius:



Therefore, we can plot it by making 3-then-2 and 2-then-3 in all directions from the center.

To find the equation, we use the two points to find it, or its square, by the *Distance Formula*:

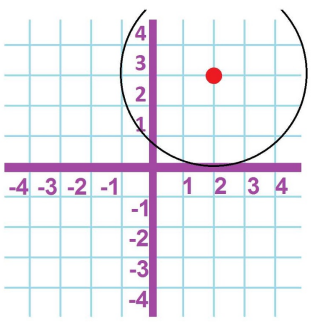
$$r^2 = 2^2 + 3^2 = 13\,.$$

Then, the equation of the circle taken from the theorem is as follows:

$$(x-2)^2 + (y-3)^2 = 13\,.$$

Example 5.6.28: circle via center and tangent

Indirect and subtler: What is the equation of the circle centered at $(2,3)$ and touching the x-axis? We need a sketch:



We realize then that the radius must be $r = 3$. Then, the equation of the circle is as follows:

$$(x-2)^2 + (y-3)^2 = 3^2\,.$$

We say that the axis is *tangent* to the circle (Chapter 2DC-3).

Exercise 5.6.29

What is the equation of the circle centered at $(1,1)$ and passing through the point $(2,3)$?

Exercise 5.6.30

What is the equation of the circle centered at $(2,3)$ and touching the y-axis?

Here is the circle we used to build trigonometric functions.

Corollary 5.6.31: Unit Circle Equation

The circle of radius 1 centered at the origin is given by the following relation:

$$x^2 + y^2 = 1$$

This fact will be used numerous times in the future.

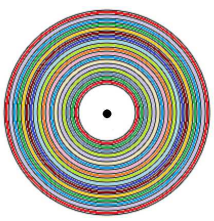
We think by analogy:

- Every parabola can be acquired from the standard one, $y = x^2$, via flips, shifts, and stretches.
- The graph of any exponential function a^x can be acquired from the standard one e^x via flips and stretches. Same for the logarithms.
- Every circle can be acquired from the standard one, $x^2 + y^2 = 1$, via shifts and stretches (flips have no effect).

First, any radius is achieved by a *uniform* stretch of the plane:

$$(x, y) \to (rx, ry)\,.$$

As we stretch the circle again and again, the "concentric" circles produced cover the whole plane:



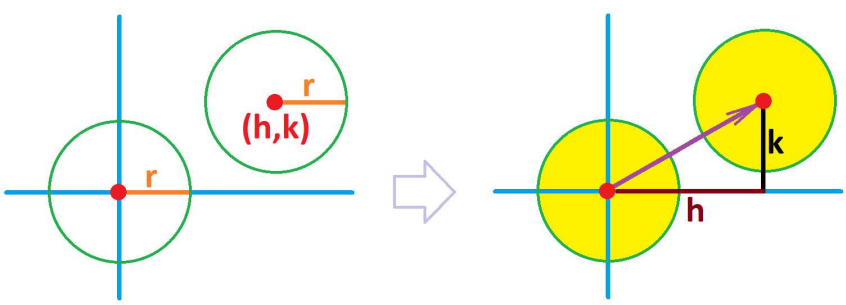
Exercise 5.6.32

Show that vertical or horizontal stretches of a parabola will produce another parabola. Show that this is not the case for circles.

Second, any location can be achieved by a shift of the plane:

$$(x, y) \to (x + h, y + k)\,.$$

We relocate our circle:



So, suppose we need the circle centered at (h, k) of radius r. We acquire it from the circle of radius 1 centered at 0 following these steps:

1. Stretch uniformly by r.
2. Shift right by h.
3. Shift up by k.

Below is the summary of this construction:

$$\begin{array}{rcrcl} x^2 & + & y^2 & = & 1 \\ (x-h)^2 & + & (y-k)^2 & = & r^2 \end{array}$$

One can still see the three transformations of the plane.

Exercise 5.6.33

Can we change the order?

Furthermore, both terms on the left are *complete squares*:

$$\begin{aligned} (x-h)^2 &= x^2 - 2xh + h^2 \\ (y-k)^2 &= y^2 - 2yk + k^2 \end{aligned}$$

To find this representation, one would need to match the terms on the right with the terms in the complete square formula:

$$(a+b)^2 = a^2 + 2ab + b^2\,.$$

Example 5.6.34: completing squares

Find the center and the radius of this circle:

$$x^2 + 2x + y^2 - 4y + 1 = 0\,.$$

Unfortunately, the equation doesn't conform to our formula! We need to manipulate the two parts of the equation towards it and find h, k, r:

$x^2 + 2x$	$y^2 - 4y$	Start with the original.
$= x^2 + 2 \cdot 1 \cdot x + 1^2 - 1^2$	$= y^2 - 2 \cdot 2 \cdot y + 2^2 - 2^2$	Add the missing term of the complete square.
$= (x^2 + 2 \cdot 1 \cdot x + 1^2) - 1$	$= (y^2 - 2 \cdot 2 \cdot y + 2^2) - 4$	Pull out the extra term.
$= (x+1)^2 - 1$	$= (y-2)^2 - 4$	Complete the square.

The terms in red are the shifts: $h = -1$ and $k = 2$.

Our equation becomes:

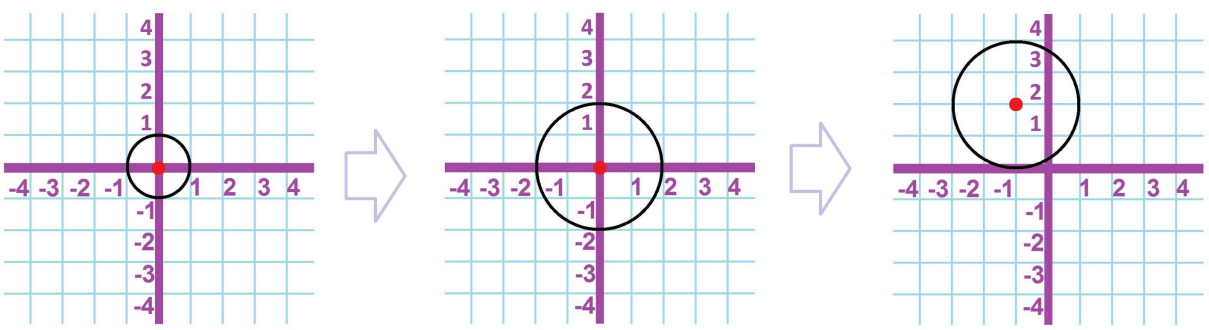
$$x^2+2x+y^2-4y+1 = (x^2+2x)+(y^2-4y)-4 = ((x+1)^2-1)+((y-2)^2-4)+1 = (x+1)^2+(y-2)^2-4\,.$$

The new form of the equation is as follows:

$$(x+1)^2 + (y-2)^2 = 2^2\,.$$

We have discovered the radius too: $r = 2$!

This is a circle centered at $(-1, 2)$ of radius 2 acquired by the following: uniform stretch by 2, shift left by 1, shift up by 2:



Exercise 5.6.35

What if we change $+1$ in the above equation to -3? What if it is -4?

Exercise 5.6.36

What are the center and the radius of this circle: $x^2 - 6x + y^2 - 4y + 12 = 0$?

Warning!

An equation of this kind can produce – in addition to a circle – a parabola, a hyperbola, and some other curves.

Theorem 5.6.37: Centered Form of Circle

Any circle on the plane can be represented in a "centered form":

$$(x-h)^2 + (y-k)^2 = r^2$$

where h, k, r are some numbers.

Exercise 5.6.38

State the converse of the theorem.

This form of a circle equation is aligned with the point-slope form of a line and the vertex form of a parabola:

	"template"	function	relation
/	line: $y = x$	$y = 2(x-1)+3$	$y-3 = 2(x-1)$ shift up by 3, shift right by 1, vertical stretch by 2
⌣	parabola: $y = x^2$	$y = 2(x-1)^2+3$	$y-3 = 2(x-1)^2$ shift up by 3, shift right by 1, vertical stretch by 2
○	circle: $x^2 + y^2 = 1$	N/A	$(y-3)^2 + (x-1)^2 = 2^2$ shift up by 3, shift right by 1, uniform stretch by 2

As you can see, these special representations reveal the transformations that have made the curve.

Exercise 5.6.39

Where is the flip in this table?

Exercise 5.6.40

Explain how the vertical stretch changes the slope of a line.

Exercise 5.6.41

What if we apply the transformation: shift up by 3, shift right by 1, vertical stretch by 2, to the unit circle?

The formula for the circle is convenient but not for plotting! The missing part in the last row of the table is explained by the fact that – unlike the other two – the circle doesn't pass the *Vertical Line Test*; it's not the graph of any function! We get around this problem in a manner similar to how we handled finding the inverse of $y = x^2$ – we split it.

We just treat our relation as an equation and solve it:

$$x^2 + y^2 = R^2$$
$$\swarrow \qquad \searrow$$
$$y = \sqrt{R^2 - x^2} \qquad y = -\sqrt{R^2 - x^2}$$

Our interpretation of the result is below.

Theorem 5.6.42: Circle as Two Graphs

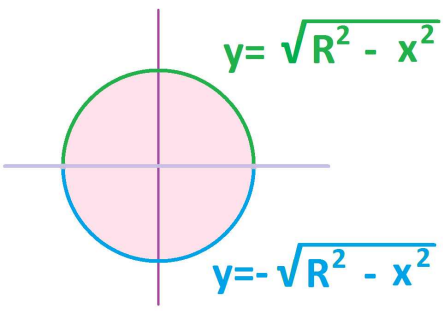
The circle of radius R centered at the origin,

$$x^2 + y^2 = R^2\,,$$

can be represented as the union of the graphs of two functions:

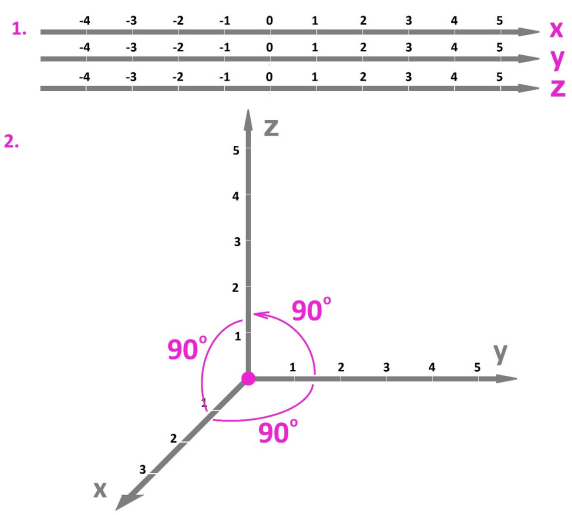
$$y = \sqrt{R^2 - x^2} \quad \text{and} \quad y = -\sqrt{R^2 - x^2}$$

We have two semicircles:



The analytic geometry of *angles*, for both dimensions, is discussed later in the chapter.

We will not address the 3-*dimensional case* (Chapter 3IC-3), but it should be obvious that the construction will start this way: "Build *three* coordinate axes and arrange them accordingly":



We then have the following geometry-algebra correspondence:

location P $\longleftrightarrow$ triple (x, y, z)

One can already see how harder it is to visualize things in the 3-dimensional space, which further justifies the need for the *algebraic* approach to geometry that we have presented in this chapter.

Collectively, these are called 1-, 2-, and 3-dimensional *Euclidean spaces*.

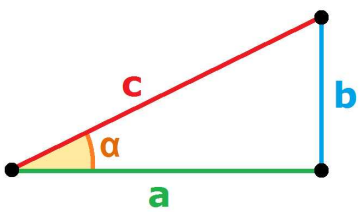
5.7. Trigonometry and the wave function

Let's review from the last chapter how the trigonometric functions emerge from *plane geometry*.

First, the *Pythagorean Theorem* states that if we have a right triangle with sides a, b, c, with c the longest one facing the right angle, then they satisfy the following:

$$a^2 + b^2 = c^2$$

Next, we concentrate on α, the angle between side a and side c:



Exercise 5.7.1

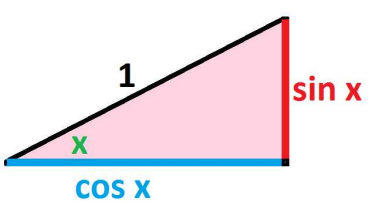
What do the ratios tell us about the trigonometric functions of the angle α?

These are the definitions:

$$\begin{aligned} \cos\alpha &= \frac{a}{c} \\ \sin\alpha &= \frac{b}{c} \end{aligned} \quad \Longrightarrow \quad \tan\alpha = \frac{b}{a} = \frac{\sin\alpha}{\cos\alpha}$$

There are many algebraic relations between the trigonometric functions. Below we examine the more important ones.

Since the trigonometric functions depend only on the proportions of the triangle, we choose to deal with the simplest kind only: a right triangle with the hypotenuse (the longest one, facing the right angle) of length 1:



The theorem then takes the following highly useful form.

Corollary 5.7.2: Pythagorean Theorem of Trigonometry

For any real number x, we have the following identity:

$$\sin^2 x + \cos^2 x = 1$$

Warning!

$\sin^2 x$ stands for $(\sin x)^2$.

Proof.

We can also derive the formula directly. We take the Pythagorean Theorem:

$$a^2 + b^2 = c^2\,,$$

and divide by c^2:

$$\frac{a^2}{c^2} + \frac{b^2}{c^2} = \frac{c^2}{c^2}\,,$$

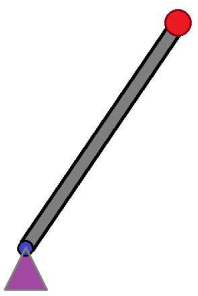
or

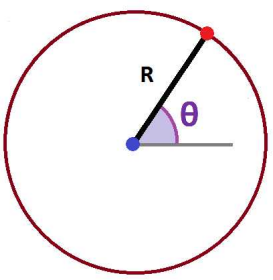
$$\left(\frac{a}{c}\right)^2 + \left(\frac{b}{c}\right)^2 = 1\,.$$

Finally, we substitute the formulas for sin and cos into this equation.

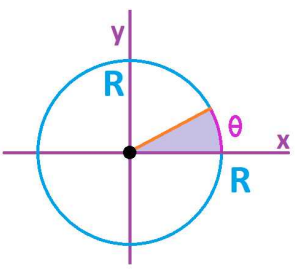
Example 5.7.3: rotating rod, parametric curve

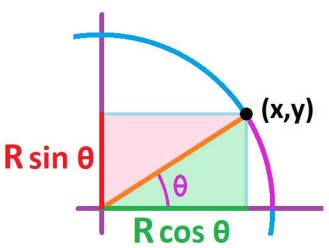
Back to our example of tracing the end of a rotating rod:

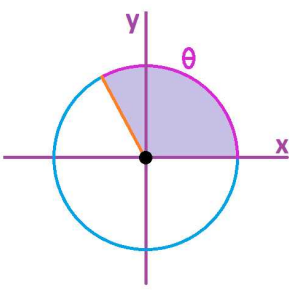




This model is where the two trigonometric functions come from:





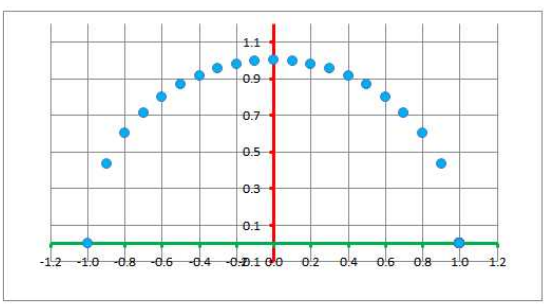


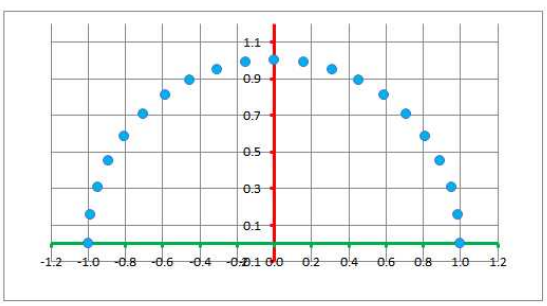
The equations, with the angle θ measured counterclockwise, give us the coordinates of points on a circle of radius R:

$$x = R\cos\theta \quad \text{and} \quad y = R\sin\theta\,.$$

We have two functions with the same input; it's a *parametric curve*.

The circle may come the graph of a function $y = f(x)$ (left) with the values of x distributed evenly:





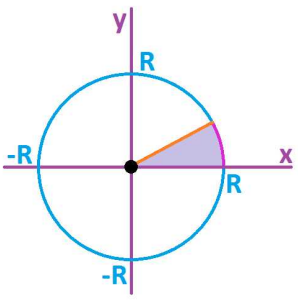
Or it may come from the two functions above (right) and this time it is not x that progresses uniformly, but the angle. The advantage of the latter representation is visible. Furthermore, with this approach nothing stops us from progressing beyond 180 degrees!

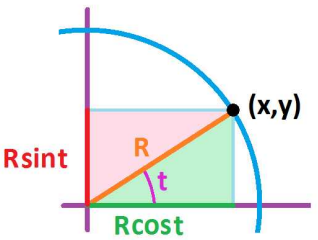
Example 5.7.4: rotating rod, continued

Alternatively, we concentrate on the actual rotation of the rod and replace the angle θ as the parameter with *time* t:

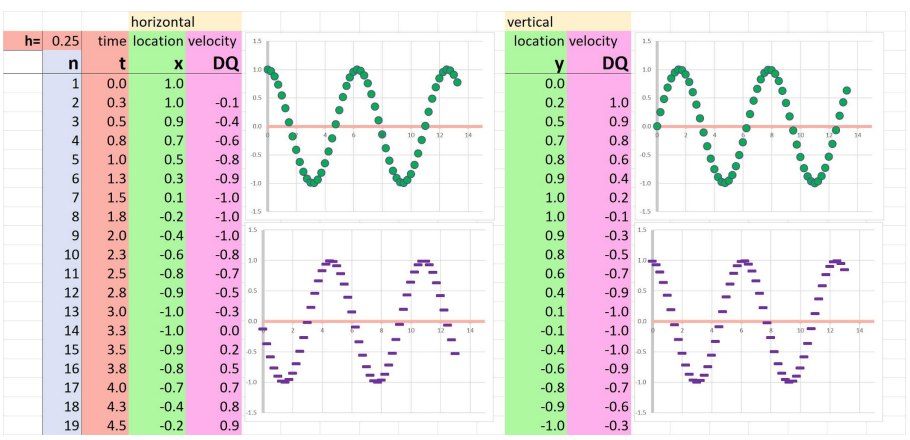
$$x = R\cos t,\ y = R\sin t\,.$$

With no restrictions on t, we have a circle traced infinitely many times:





Let's make it specific, $R = 1$. There are two functions: the horizontal and the vertical locations, possibly shadows on the ground and a wall. We plot them below, x against t and y against t (top row):



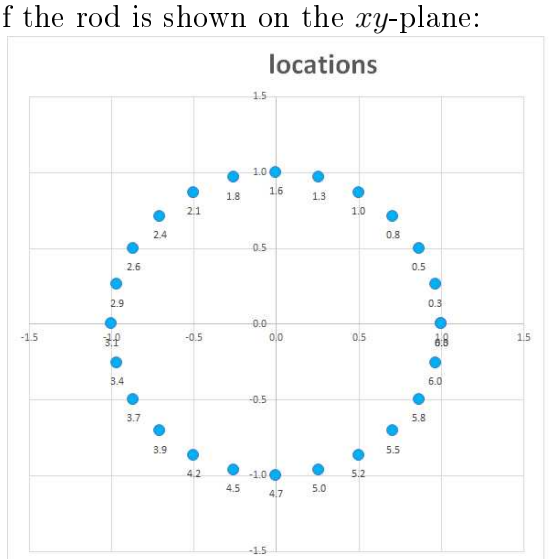
		horizontal			vertical	
h=	0.25	time	location	velocity	location	velocity
	n	t	x	DQ	y	DQ
	1	0.0	1.0		0.0	
	2	0.3	1.0	-0.1	0.2	1.0
	3	0.5	0.9	-0.4	0.5	0.9
	4	0.8	0.7	-0.6	0.7	0.8
	5	1.0	0.5	-0.8	0.8	0.6
	6	1.3	0.3	-0.9	0.9	0.4
	7	1.5	0.1	-1.0	1.0	0.2
	8	1.8	-0.2	-1.0	1.0	-0.1
	9	2.0	-0.4	-1.0	0.9	-0.3
	10	2.3	-0.6	-0.8	0.8	-0.5
	11	2.5	-0.8	-0.7	0.6	-0.7
	12	2.8	-0.9	-0.5	0.4	-0.9
	13	3.0	-1.0	-0.3	0.1	-1.0
	14	3.3	-1.0	0.0	-0.1	-1.0
	15	3.5	-0.9	0.2	-0.4	-1.0
	16	3.8	-0.8	0.5	-0.6	-0.9
	17	4.0	-0.7	0.7	-0.8	-0.7
	18	4.3	-0.4	0.8	-0.9	-0.6
	19	4.5	-0.2	0.9	-1.0	-0.3

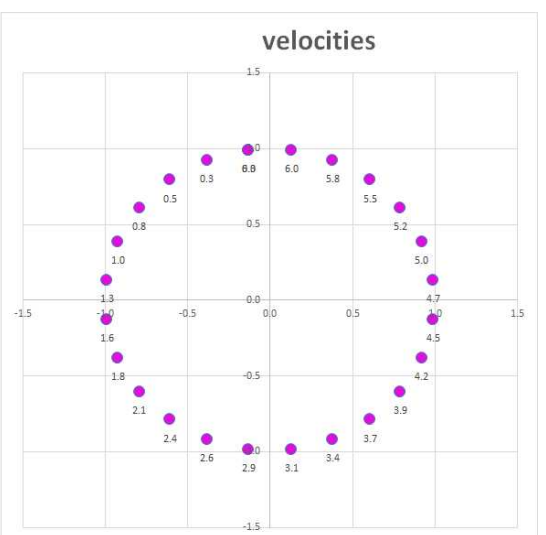
f

$\frac{\Delta f}{\Delta t}$

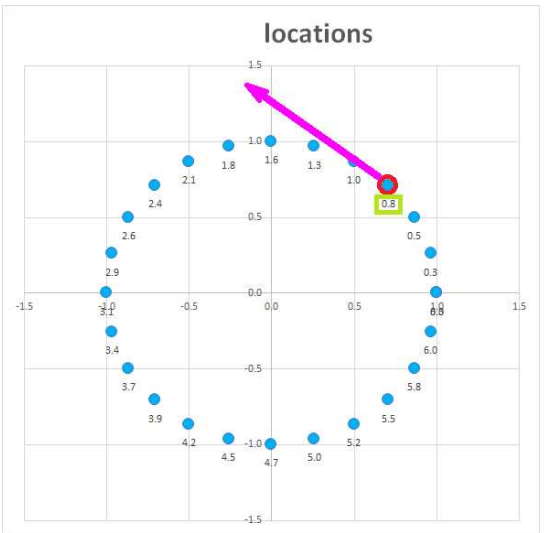
Furthermore, either function has its difference quotient, i.e., the horizontal and the vertical velocities (bottom row). If these two are changing disproportionally, we should be able to see a change in *direction* if we combine them.

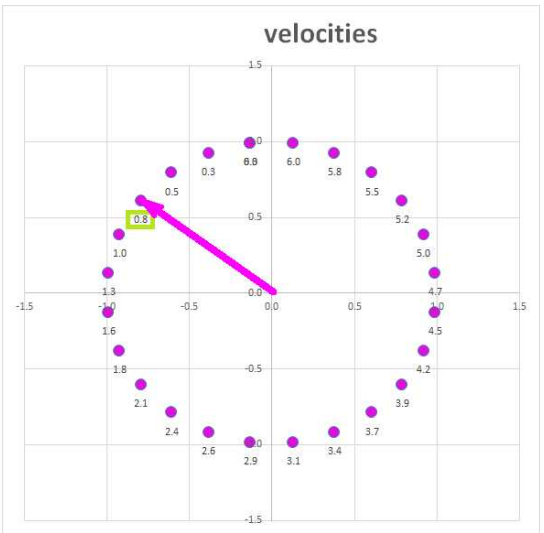
So, two different observers will see two different side-to-side movements. However, the *trajectory* of the end of the rod is shown on the xy-plane:





There is no t-axis but the locations and the velocities are marked accordingly. By matching these time stamps we can find the velocity for each location:





This arrow found on the right is then attached to the location on the left. This is the direction of motion at this moment!

Exercise 5.7.5

(a) When is the horizontal velocity the highest? (b) Where it the horizontal velocity the highest? (c) What pattern do the velocities exhibit?

Exercise 5.7.6

Answer the questions in the last exercise for these two alternative ways to move along this circle: (a) backward direction:

$$x = \cos(-t),\ y = \sin(-t)\,;$$

and (b) twice as fast:

$$x = \cos 2t,\ y = \sin 2t\,.$$

Exercise 5.7.7

Explain the plot of the velocities.

Exercise 5.7.8

Design a parametric curve (or two) for the clock.

Example 5.7.9: values of sin

When the angle is 45 degrees, the sine and cosine are equal:

$$\sin\frac{\pi}{4} = \cos\frac{\pi}{4}\,.$$

Therefore, the *Pythagorean Theorem* implies that

$$\sin\frac{\pi}{4} = \cos\frac{\pi}{4} = \frac{\sqrt{2}}{2}\,.$$

Exercise 5.7.10

Find all possible values of x ("preimage") for which

$$\sin x = \frac{\sqrt{2}}{2}\,.$$

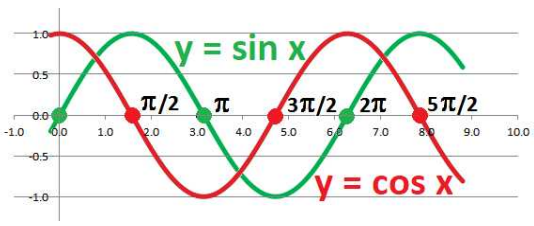
Exercise 5.7.11

Derive these formulas:

$$\sin\theta = \pm\frac{\tan\theta}{\sqrt{1+\tan^2\theta}};$$
$$\cos\theta = \pm\frac{1}{\sqrt{1+\tan^2\theta}}.$$

The Pythagorean Theorem above establishes are relation between the two main trigonometric functions, $\sin x$ and $\cos x$; if we know one, we know the other. Are there other such relations?

The two functions are 2π-periodic, which means that when shifted horizontally by 2π, the graph lands exactly on itself:

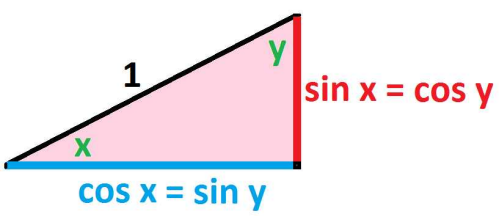


But they also appear to be shifts of each other! Let's confirm this observation.

We go back to the definition. If the hypotenuse is 1, the definition of the sine and the cosine of an angle simplifies:

$$\cos(\text{ angle }) = \text{adjacent side}$$
$$\sin(\text{ angle }) = \text{opposite side}$$

But this definition applies to either of the two other angles of our triangle, x an y:



As x is replaced with y, adjacent becomes opposite and vice versa. We have, consequently:

$$\sin x = \cos y \quad \text{and} \quad \cos x = \sin y\,.$$

Furthermore, since x and y are the two angles of the right triangle, they add up to 90 degrees:

$$x + y = \pi/2\,.$$

Then, we substitute $y = \pi/2 - x$ into the two formulas proving the following.

Theorem 5.7.12: Sine and Cosine of Complementary Angles

For any x, we have the following identities:

$$\sin x = \cos(\pi/2 - x) \quad \text{and} \quad \cos x = \sin(\pi/2 - x)$$

Therefore, the graph of one function is the graph of the other horizontally flipped and then shifted right by $\pi/2$.

Exercise 5.7.13

Explain the last statement.

A most important conclusion is as follows:

▶ The shapes of graphs of sine and cosine are identical.

However, an even closer relation is visible in the graph: The graph of the cosine shifted right by $\pi/2$ is the graph of the sine. To prove this, we just take the formula in the theorem and use the fact that the cosine is even:

$$\sin x = \cos(\pi/2 - x) = \cos(-(\pi/2 - x)) = \cos(x - \pi/2)\,.$$

These are the formulas:

Corollary 5.7.14: Shift of Sine is Cosine

For any x, we have the following identities:

$$\sin x = \cos(x - \pi/2) \quad \textit{and} \quad \cos x = \sin(x + \pi/2)$$

The following are two important trigonometric formulas that allow us to represent the sine and the cosine of the sum of two angles in terms of the two trigonometric functions of the two angles.

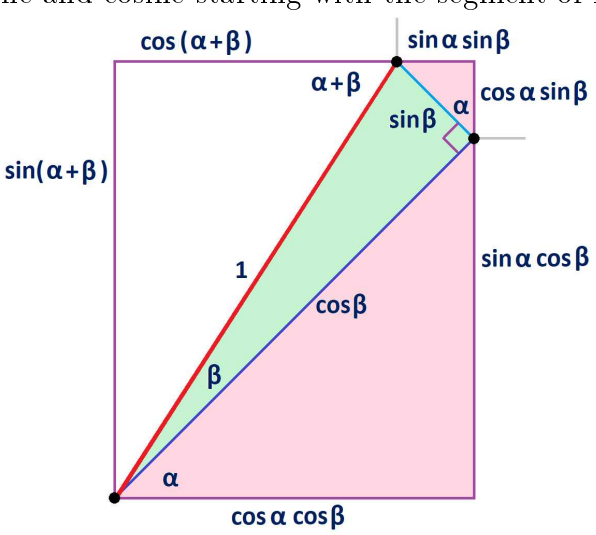
Theorem 5.7.15: Sine and Cosine of Sum

For any two angles α and β, we have the following identities:

$$\begin{aligned} \sin(\alpha+\beta) &= \sin\alpha\cos\beta + \cos\alpha\sin\beta \\ \cos(\alpha+\beta) &= \cos\alpha\cos\beta - \sin\alpha\sin\beta \end{aligned}$$

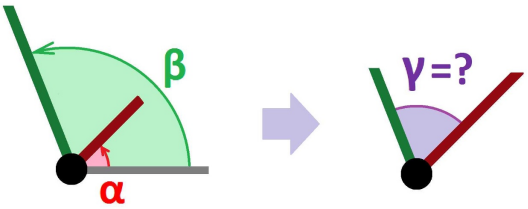
Proof.

The proof is for the acute α, β, $\alpha+\beta$. The labels for the sides of the (right) triangles are deduced from the definitions of the sine and cosine starting with the segment of length 1 (red):



The fact that the horizontal sides of the rectangle are equal produces the former formula, and the fact that the vertical sides of the rectangle are equal produces the latter.

Later in this chapter (and then in Chapter 2DC-3), we will need a trigonometric formula that allows us to represent the cosine of the difference of two angles in term of the two trigonometric functions of the two angles:



Corollary 5.7.16: Sine and Cosine of Difference

For any two angles α and β, we have the following identities:

$$\begin{aligned}\sin(\alpha-\beta) &= \sin\alpha\cos\beta - \cos\alpha\sin\beta\\ \cos(\alpha-\beta) &= \cos\alpha\cos\beta + \sin\alpha\sin\beta\end{aligned}$$

Proof.

To derive the result from the theorem, we simply use the oddness of the sine and the evenness of the cosine.

Corollary 5.7.17: Sine and Cosine of Double Angle

For any angle θ, we have the following identities:

$$\begin{aligned}\sin 2\theta &= 2\sin\theta\cos\theta\\ \cos 2\theta &= \cos^2\theta - \sin^2\theta = 1 - 2\sin^2\theta = 2\cos^2\theta - 1\end{aligned}$$

Proof.

Just choosing $\alpha = \beta = \theta$ in the theorem (and using the Pythagorean Theorem for the second part) produces the formula.

Corollary 5.7.18: Sine and Cosine of Half-Angle

For any angle α, we have the following identities:

$$\sin\left(\frac{\alpha}{2}\right) = \pm\sqrt{\frac{1-\cos\alpha}{2}} \quad \textit{and} \quad \cos\left(\frac{\alpha}{2}\right) = \pm\sqrt{\frac{1+\cos\alpha}{2}}$$

Proof.

From the second identity in the corollary, we take:

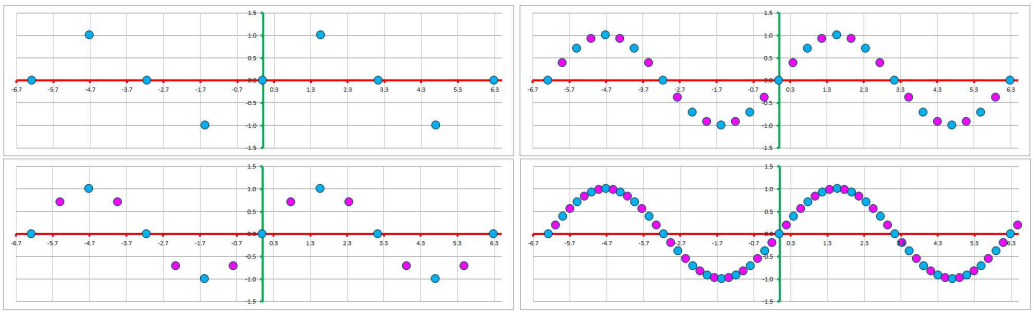
$$\cos 2\theta = 1 - 2\sin^2\theta\,,$$

set $\alpha = 2\theta$, and solve, producing the first formula. From the second identity, we also take:

$$\cos 2\theta = 2\cos^2\theta - 1\,,$$

set $\alpha = 2\theta$, and solve, producing the second formula.

The formula below allowed us in the last chapter to compute the values of the sine and cosine for denser and denser values of x:



Corollary 5.7.19: Sine and Cosine of Average Angle

For any angles α and β, we have the following identities:

$$\begin{aligned}\sin\left(\frac{\alpha+\beta}{2}\right) &= \pm\sqrt{\frac{1-\cos\alpha\cos\beta-\sin\alpha\sin\beta}{2}}\\ \cos\left(\frac{\alpha+\beta}{2}\right) &= \pm\sqrt{\frac{1+\cos\alpha\cos\beta-\sin\alpha\sin\beta}{2}}\end{aligned}$$

Proof.

We simply combine the last theorem and the last corollary.

We will use the following in this section:

Corollary 5.7.20: Sum of Sines and Cosines

For any α and β, we have the following identities:

$$\begin{aligned}\sin\alpha+\sin\beta &= 2\sin\left(\frac{\alpha+\beta}{2}\right)\sin\left(\frac{\alpha-\beta}{2}\right)\\ \cos\alpha+\cos\beta &= 2\cos\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right)\end{aligned}$$

Proof.

Let

$$A=\alpha+\beta,\ B=\alpha-\beta\,.$$

Therefore:

$$\alpha=(A+B)/2,\ \beta=(A-B)/2\,.$$

We need to simplify this:

$$\sin\alpha+\sin\beta=\sin((A+B)/2)+\sin((A-B)/2)\,.$$

We apply the theorem to either part:

$$\begin{aligned}\sin((A+B)/2) &= \sin A/2\cos B/2 &&+\cos A/2\sin B/2\,;\\ \sin((A-B)/2) &= \sin A/2\cos B/2 &&-\cos A/2\sin B/2\,.\end{aligned}$$

Adding these two, we find:

$$\sin(A+B)/2+\sin(A-B)/2)=2\sin A/2\cos B/2\,.$$

The formulas follow.

Exercise 5.7.21

Provide detailed proofs of these results.

Exercise 5.7.22

Solve the equation below:

$$\cos^2(x^2+1) = .2\,.$$

Make up your own equation and solve it. Repeat.

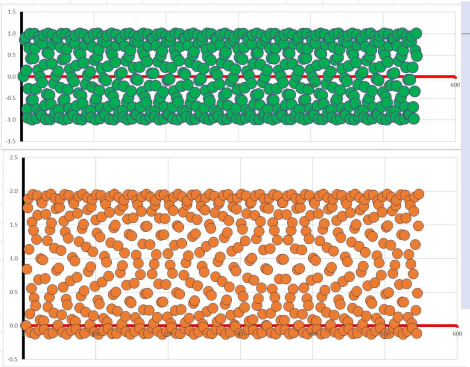
Example 5.7.23: sequence $\sin n$

Consider the sequence

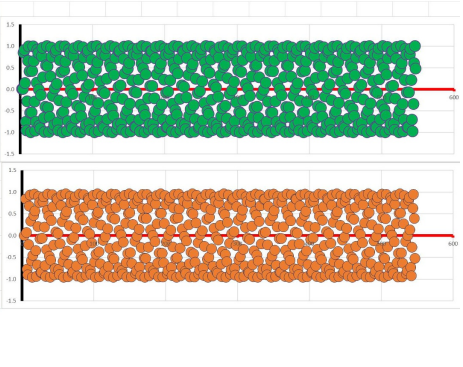
$$a_n = \sin n\,.$$

Because the increment of the input isn't proportional to the period of the function, it gives an appearance of randomness. The values look uniformly spread (between -1 and 1):

n	a_n	sum
0		0.00
1	0.84	0.84
2	0.91	1.75
3	0.14	1.89
4	-0.76	1.14
5	-0.96	0.18
6	-0.28	-0.10
7	0.66	0.55
8	0.99	1.54
9	0.41	1.96
10	-0.54	1.41
11	-1.00	0.41
12	-0.54	-0.13
13	0.42	0.29
14	0.99	1.29
15	0.65	1.94
16	-0.29	1.65
17	-0.96	0.69
18	-0.75	-0.06



n	a_n	diff
0	0.00	
1	0.84	0.84
2	0.91	0.07
3	0.14	-0.77
4	-0.76	-0.90
5	-0.96	-0.20
6	-0.28	0.68
7	0.66	0.94
8	0.99	0.33
9	0.41	-0.58
10	-0.54	-0.96
11	-1.00	-0.46
12	-0.54	0.46
13	0.42	0.96
14	0.99	0.57



Then the values of its sum Σa_n and difference Δa_n (Chapter 1) also seem restricted to a certain interval!

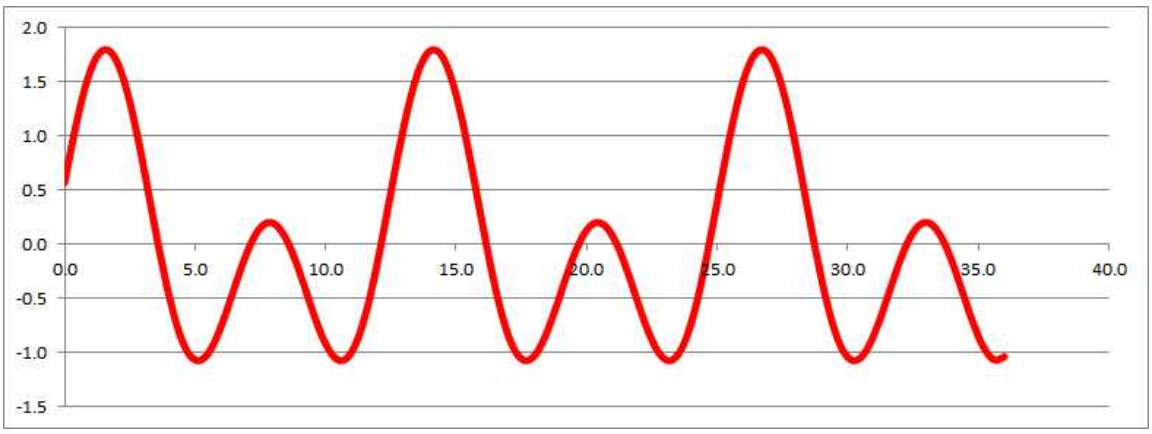
Exercise 5.7.24

Prove the last statement using the above results.

Recall how in Chapter 3 we produced the graphs of *all* quadratic functions as transformations of one parabola: the Vertex Formula of Quadratic Polynomial. Based on the above results, we also have a common name for all transformations of the graph of $y = \sin x$, the *sinusoid*:

template	general	name of the curve
$y = x^2$	$y = a(x-h)^2 + k$	*parabola*
$y = \sin x$	$y = A\sin(\omega x + \phi)$	*sinusoid*
$y = \sin x$	$y = B\cos(\omega x + \phi)$	*sinusoid*

We will learn how to use the trigonometric functions to represent numerous kinds of periodic phenomena. These phenomena are often represented by a single quantity that changes in a repetitive manner:



The simplest functions of this kind are trigonometric.

Definition 5.7.25: wave function

The function

$$f(x) = A\sin(\omega x + \phi)$$

is called a *wave function*. Furthermore,

- $A > 0$ is called the *amplitude*,
- $\omega > 0$ is called the *frequency*, and
- ϕ is called the *phase*,

of the wave function.

These three parameters change the original graph of $y = \sin x$ in predictable ways:

amplitude	A	$y \to Ay$	vertical stretch
frequency	ω	$x \to \omega x$	horizontal shrink
phase	ϕ	$x \to x + \phi$	horizontal shift

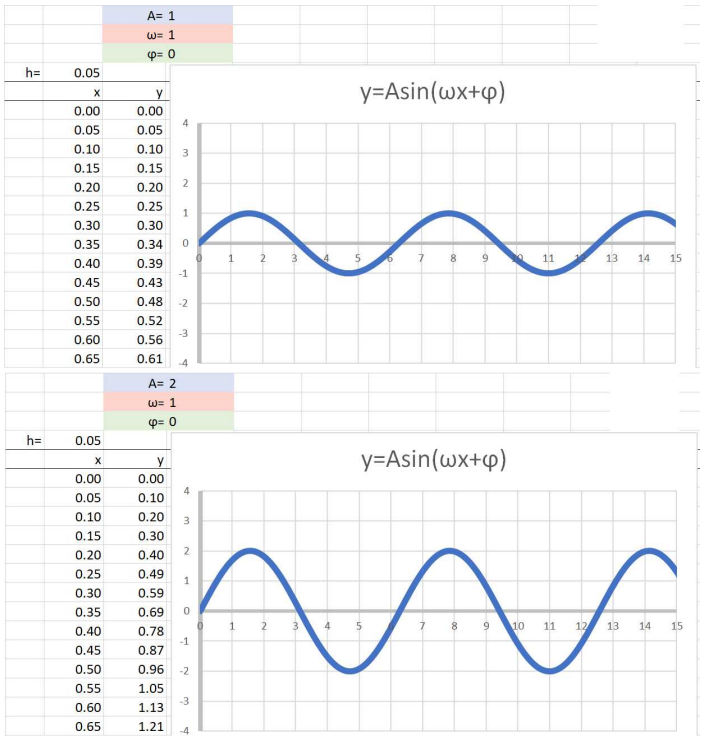
We place the parameters in a spreadsheet and plot the function computed with the following formula for y:

```
=R1C[1]*SIN(R2C[1]*RC[-1]+R3C[1])
```

We vary these parameters below.

Example 5.7.26: wave function, amplitude varies

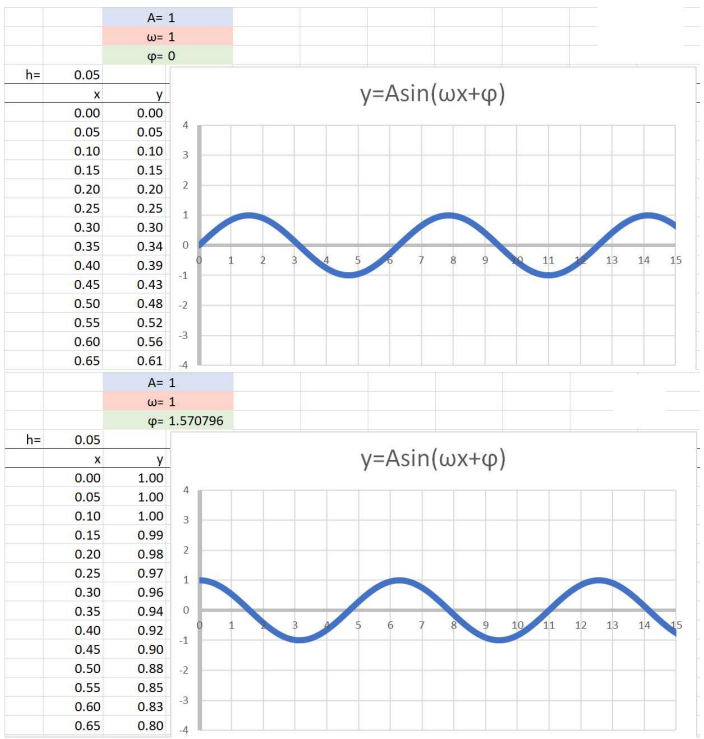
As the values of the sine run between -1 and 1, the values of this function run between $-A$ and A. It's a vertical stretch/shrink:



A stretch by a factor of 2 is shown. We can see that the function still repeats itself every 2π; just the swing has doubled.

Example 5.7.27: wave function, phase varies

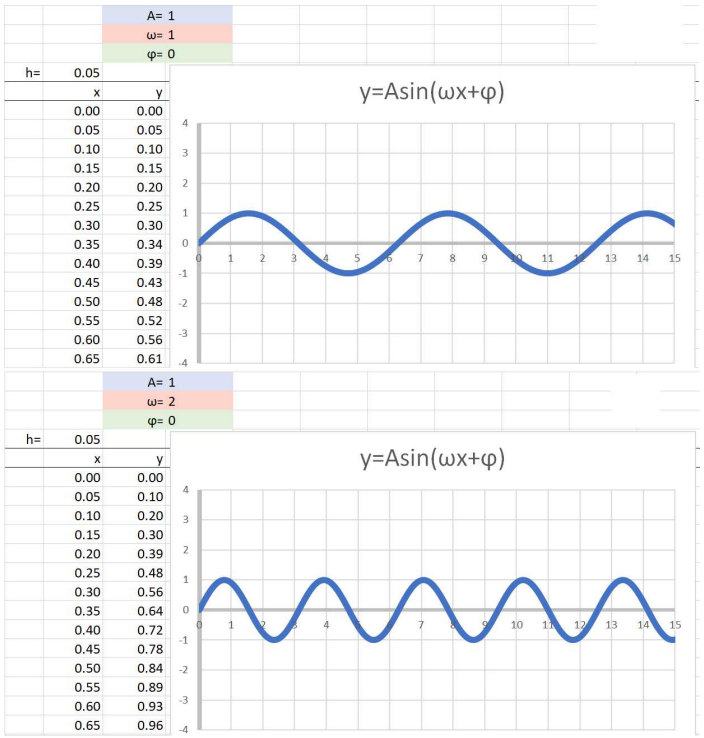
The phase ϕ is just a horizontal shift:



A shift of $\pi/2$ right is shown. This is the cosine!

Example 5.7.28: wave function, frequency varies

As the sine is 2π-periodic, this function's period is $2\pi/\omega$. It's a horizontal stretch/shrink:



A shrink by a factor of 2 is shown. We can see that the function repeats itself *twice* within every interval of 2π; its period is $2\pi/2 = \pi$.

Exercise 5.7.29

Prove the statement about the periods.

The common name "sinusoid" is justified by the fact that all these curves look the same regardless of the choice of the parameters.

What if we *add* two wave functions with different values of the same parameter? Such an addition corresponds to, for example, two sounds heard at the same time.

We know how to add functions, but we would like to predict the kind of function will result:

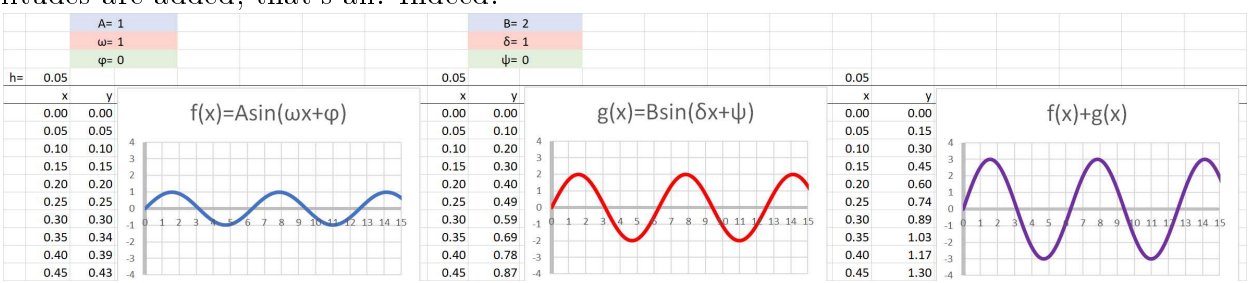
$$\begin{array}{ll} f(x) & = A\sin(\omega x + \phi) \\ g(x) & = B\sin(\delta x + \psi) \\ \hline f(x)+g(x) & = \,? \end{array}$$

Example 5.7.30: adding wave functions, different amplitudes

What if we combine two wave functions with different amplitudes? Consider:

$$y = \sin x + 2\sin x = 3\sin x\,.$$

Amplitudes are added, that's all! Indeed:



Exercise 5.7.31

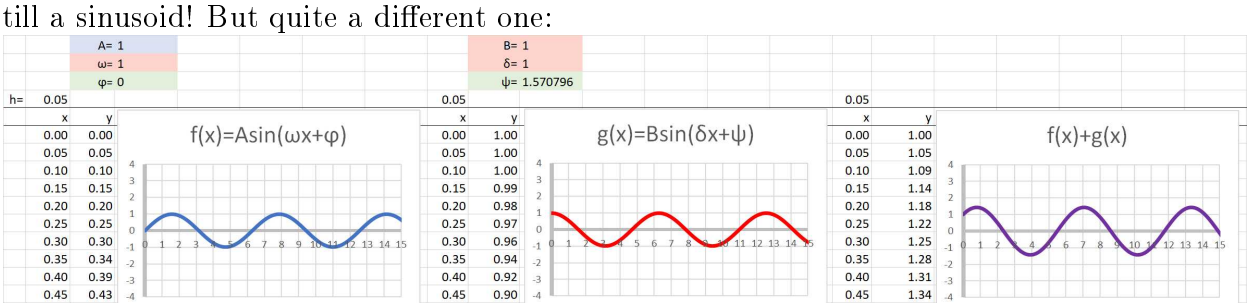
Make a general statement about adding wave functions with different amplitudes and prove it.

Example 5.7.32: adding wave functions, different phases

What if we combine two wave functions with different phases? Consider:

$$\begin{aligned} y &= \sin(x) + \sin(x + \pi/2) \\ &= 2\sin\left(\frac{x + (x+\pi/2)}{2}\right)\sin\left(\frac{x - (x+\pi/2)}{2}\right), \quad \text{by Sum of Sines and Cosines formula} \\ &= 2\sin\left(\frac{2x+\pi/2}{2}\right)\sin\left(\frac{-\pi/2)}{2}\right) \\ &= -\sqrt{2}\sin\left(x + \pi/4\right) . \end{aligned}$$

It's still a sinusoid! But quite a different one:



Exercise 5.7.33

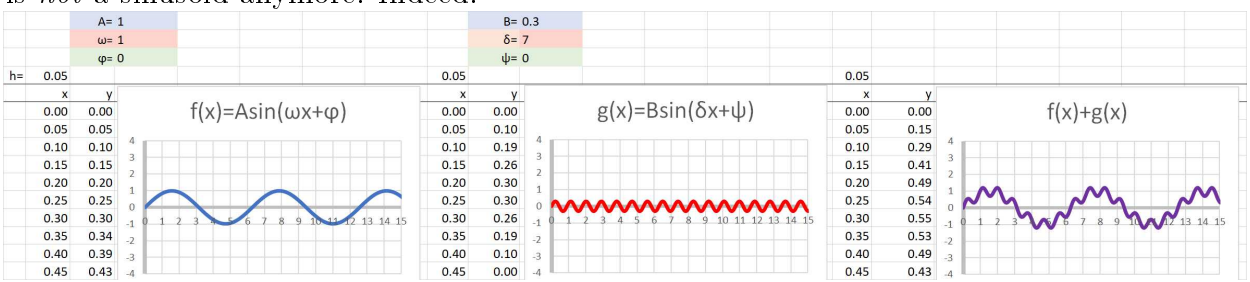
Make a general statement about adding wave functions with different phases and prove it.

Example 5.7.34: adding wave functions, different frequencies

What if we combine two wave functions with different frequencies? Consider:

$$y = \sin(x) + .3\sin(7x) .$$

This is *not* a sinusoid anymore! Indeed:



It is still 2π-periodic! Also, highlighted by the difference in amplitudes, both of the frequencies are clearly visible. If the wave functions represent sound, the second function g may be the *noise*. It is then our challenge to go in the backward direction, i.e., from $f + g$ to f and g, in order to remove it.

Exercise 5.7.35

Show that the sum of two wave functions $\sin x + \sin 3x$ has period 2π.

Exercise 5.7.36

What happens when the frequencies aren't proportional, such as $\sin x + \sin \pi x$?

We now apply this, time-dependent, interpretation of periodic motion to the rotation problem. There will be *two wave functions* producing a parametric representation of the circle.

We start with the angle-dependent representation of the circle of radius R centered at 0:

$$\begin{cases} x &= R\cos\theta\,, \\ y &= R\sin\theta\,. \end{cases}$$

But in the new model the angle, θ, depends on time. We assume a constant speed of rotation (angular velocity) and, therefore, a constant frequency :

$$\theta = \omega t + \phi\,.$$

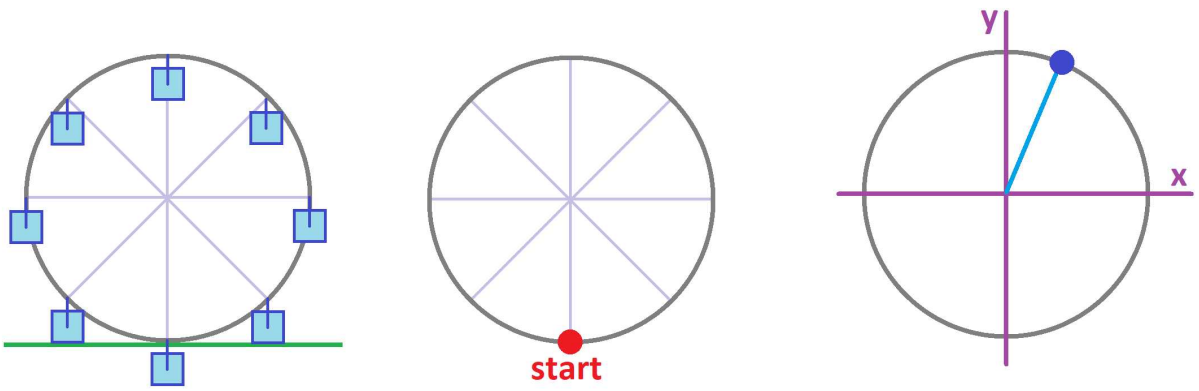
Substitution produces this parametric curve:

$$\begin{cases} x &= R\cos(\omega t + \phi)\,, \\ y &= R\sin(\omega t + \phi)\,. \end{cases}$$

Example 5.7.37: Ferris wheel

Let's consider a specific case of the Ferris wheel:

- It has a radius of 100 feet and
- It makes a full turn in 2 minutes.



We then can ask questions about *where* (the first two) as well as about *when* (third):

1. How far away from the base, horizontally, are you in 5 minutes?
2. How high are you in 20 seconds?
3. When are you 30 feet above the ground?

First, we need a complete model; we need to find the *three* parameters for the two functions above. Of course, the radius is

$$R = 100\,.$$

Second, making a full turn (2π radians) in 2 minutes means that the frequency is

$$\omega = \frac{2\pi}{2} = \pi\,.$$

Third, the starting point of the trip is at the bottom of the wheel, i.e., with the rotating segment pointing down; therefore, the phase is

$$\phi = -\pi/2\,.$$

So, the model's equations make up the following *parametric curve* (time in minutes):

$$\begin{cases} x &= 100\cos(\pi t - \pi/2)\,, \\ y &= 100\sin(\pi t - \pi/2)\,. \end{cases}$$

They can answer all questions we may have.

For the first question, we just substitute $t = 5$ into the first equation:

$$x = 100\cos(\pi\cdot 5 - \pi/2) = 100\cos(4\pi + \pi/2) = 100\cos(\pi/2) = 100\cdot 0 = 0 \text{ feet.}$$

That's how high we are at that time.

For the second question, we substitute $t = 1/3$ (i.e., 20 seconds expressed in minutes) into the second equation:

$$y = 100 \sin\left(\pi \cdot \frac{1}{3} - \frac{\pi}{2}\right) = 100 \sin\left(-\frac{\pi}{6}\right) = 100 \cdot (-.5) = -50 \text{ feet.}$$

That's $100 - 50 = 50$ feet above the ground.

For the third question, we substitute $y = -100 + 30 = -70$ into the second equation, and solve:

$$\begin{array}{lll} -70 & = 100 \sin(\pi t - \pi/2) & \Longrightarrow \\ \sin(\pi t - \pi/2) & = -.7 & \Longrightarrow \\ \pi(t - 1/2) & = \arcsin(-.7) & \Longrightarrow \\ t & = \dfrac{1}{\pi} \arcsin(-.7) + \dfrac{1}{2} & \\ & \approx .25 \text{ minutes.} & \end{array}$$

That's the time when we reach this height for the first time. (We will continue our study of parametric curves in Chapter 3IC-4.)

Exercise 5.7.38

Finish the answer to the third question.

We have seen a lot of trigonometric functions. In spite of the wide applicability of the *Sine of Sum* formula and its corollaries, only these are important enough to memorize:

Trigonometric Identities

Descriptions:	Formulas:	
Sine and cosine are periodic:	$\sin(x + 2\pi) = \sin(x)$,	$\cos(x + 2\pi) = \cos(x)$
Sine is odd and cosine even:	$\sin(-x) = -\sin(x)$,	$\cos(-x) = \cos(x)$
Sine and cosine are shifts of each other:	$\sin(x + \pi/2) = \cos(x)$,	$\cos(x - \pi/2) = \sin(x)$
Pythagorean Theorem:	$\sin^2 x + \cos^2 x = 1$	

5.8. The Euclidean plane: angles

In this section, we take up the second geometric task, evaluating *directions*, in the Euclidean space equipped with the Cartesian coordinate system.

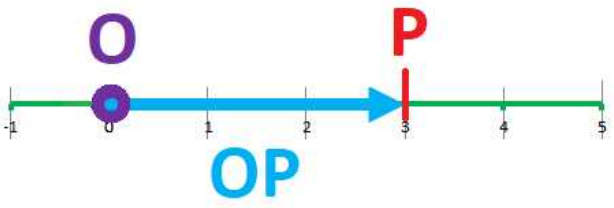
First, the 1-*dimensional* Euclidean space, i.e., nothing but the x-axis.

What does a *direction* on the real line mean? We will pursue the following approach.

Definition 5.8.1: vector in dimension 1

If a line segment's starting point is the origin, i.e., it's OP for some point $P \neq O$, it is called a (1-dimensional) *vector* in $\mathbf{R}$.

Vectors are usually visualized as arrows:



Vectors have two main attributes:

- A vector has its *magnitude*, which is the absolute value of the coordinate x of its terminal point: $|x|$.
- A vector has its *direction*, which is, in the one-dimensional case, either positive or negative.

Now, how do we *compare* the directions of two vectors?

Suppose we have two points: $P \neq O$, $Q \neq O$. We deal with the directions from the origin O toward locations P and Q. Of course, there can be only two outcomes:

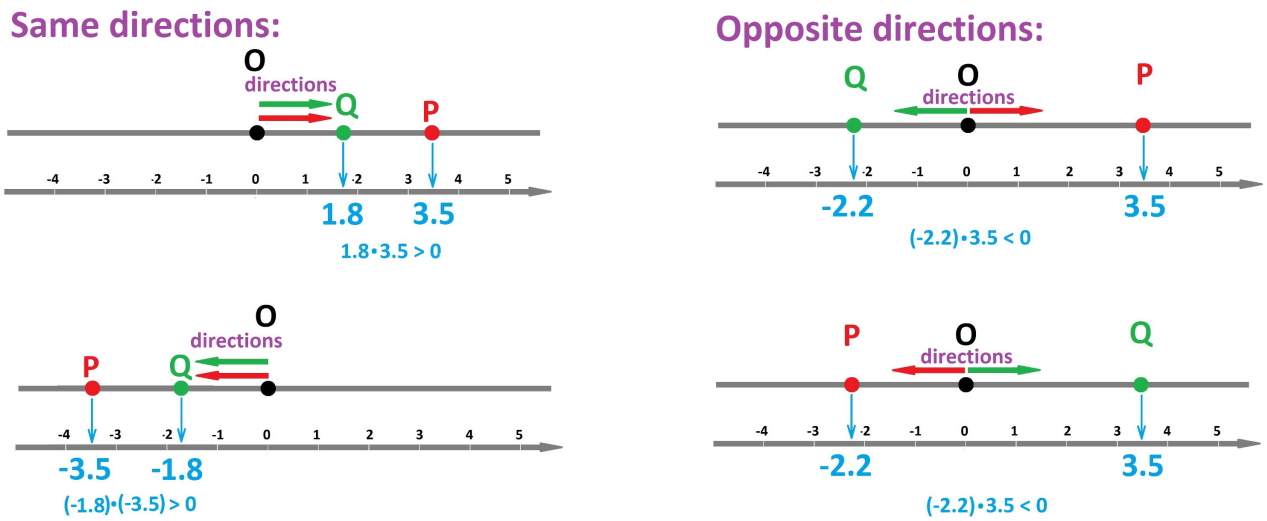
- If P and Q are on the same side of O, then the *directions are same*:

$$P\ Q\ \leftarrow O \rightarrow \quad \text{or} \quad \leftarrow O \rightarrow\ P\ Q$$

- If P and Q are on the opposite sides of O, then the *directions are opposite*:

$$P\ \leftarrow O \rightarrow\ Q \quad \text{or} \quad Q\ \leftarrow O \rightarrow\ P$$

Let's examine the coordinates. These are the four possibilities:



When the two vectors are represented by their coordinates, x and x', the analysis of their directions becomes algebraic:

- If $x > 0,\ x' > 0$ or $x < 0,\ x' < 0$, then the directions are the same.
- If $x > 0,\ x' < 0$ or $x < 0,\ x' > 0$, then the directions are the opposite.

Fortunately, the *product* provides us with a single expression that makes this determination:

points:	$P\ Q\ \leftarrow O \rightarrow$	$\leftarrow O \rightarrow\ P\ Q$	same
signs:	$- \cdot - = +$	$+ \cdot + = +$	$\boxed{+}$
points:	$P\ \leftarrow O \rightarrow Q$	$Q\ \leftarrow O \rightarrow\ P$	opposite
signs:	$- \cdot + = -$	$- \cdot + = -$	$\boxed{-}$

Theorem 5.8.2: Directions for Dimension 1

The directions from 0 *to* $x \neq 0$ *and* $x' \neq 0$ *are*

- *the same when* $x \cdot x' > 0$*; and*
- *the opposite when* $x \cdot x' < 0$*.*

Let's restate the theorem in terms of the sign function:

sign of the product	directions		angle
$\text{sign}(x \cdot x') = 1$	same	$\overrightarrow{\ }\overrightarrow{\ }$	0 degrees
$\text{sign}(x \cdot x') = -1$	opposite	$\overrightarrow{\ }\overleftarrow{\ }$	180 degrees

We also measure the actual angles between the vectors (last column).

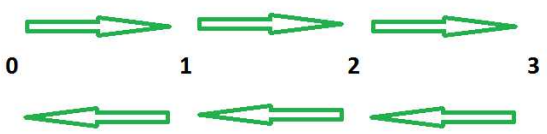
Exercise 5.8.3

Where else did we see the relation: $1 \leftrightarrow$ 0 degrees and $-1 \leftrightarrow$ 180 degrees?

This brings us to the idea of an *oriented interval*, or segment, within the axis. Any two distinct points P and Q produce *two* intervals: PQ and QP. These two intervals have opposite directions, and we say that they have opposite *orientations*. Furthermore, since the direction of the x-axis is set ahead of time, we can match the direction of the interval with it, as follows.

Definition 5.8.4: positively and negatively oriented segments

- If $P < Q$, the segment PQ is called *positively oriented*.
- If $P > Q$, the segment PQ is called *negatively oriented*.

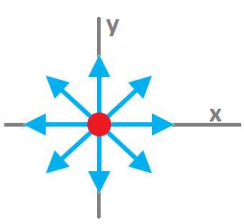


In fact, we adopt the following convention:

$$QP = -PQ\,.$$

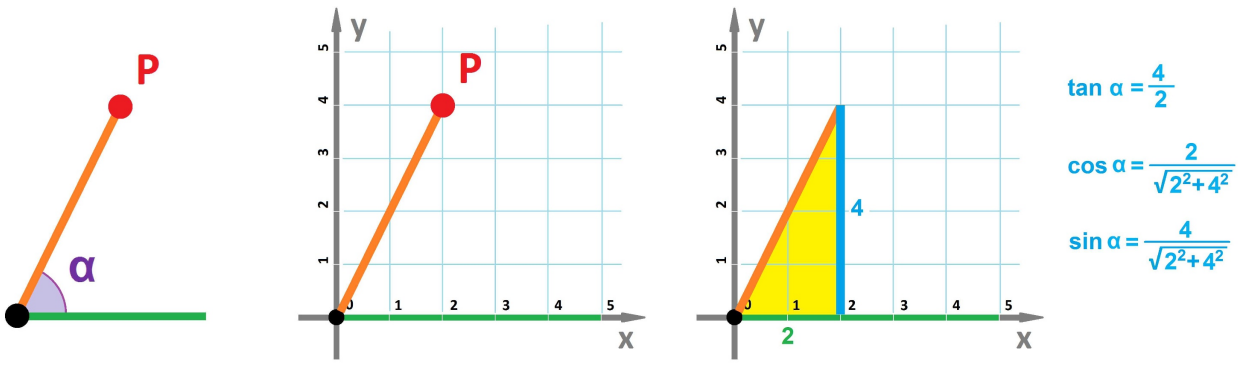
So, a ruler has *two* ends and there are *two* ways to place it along a measuring tape.

Second, the 2-*dimensional* Euclidean space, i.e., the xy-plane. There are infinitely many directions now:



The issue of the direction of a single line (or a vector) has been solved: It's the slope! Let's review.

The question is the one about the angle between the line OP with the x-axis. It is determined from the coordinates of $P = (x, y)$ via this simple trigonometry:



Exercise 5.8.5

Explain these computations.

One of these formulas has been especially important.

Theorem 5.8.6: Slope is Tangent

The tangent of the angle α between the x-axis and the line from O to a point $P = (x, y) \neq O$ is equal to the slope of this line:

$$\tan \alpha = \frac{y}{x}$$

These formulas will be used later.

Theorem 5.8.7: Sine and Cosine of Direction

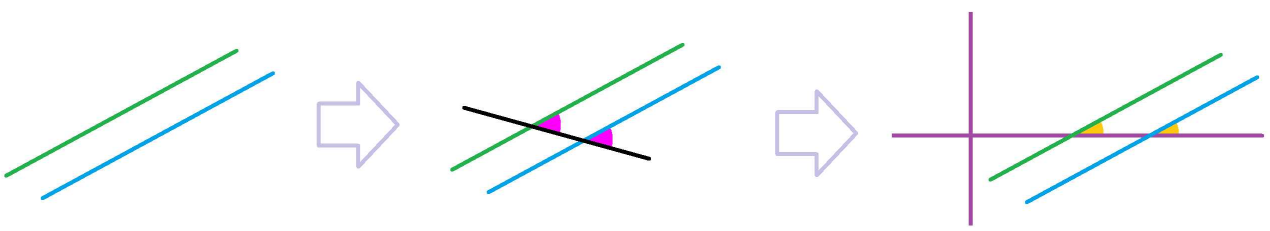
The sine and the cosine of the angle α between the x-axis and the line from O to a point $P = (x, y) \neq O$ are given by:

$$\cos \alpha = \frac{x}{\sqrt{x^2 + y^2}}$$
$$\sin \alpha = \frac{y}{\sqrt{x^2 + y^2}}$$

Exercise 5.8.8

Prove the theorem.

Recall from Euclidean geometry that two lines are called *parallel* if they form the same angle with a given line (middle):



If we choose this other line to be the x-axis (right), we can apply the above theorem to the angle α. We conclude the following:

Theorem 5.8.9: Parallel Lines, Same Slope

Two (non-vertical) lines on the xy-plane are parallel if and only if they have equal slopes.

The statement can be abbreviated as follows:

$$y = mx + b \;||\; y = m'x + b' \iff m = m'\,.$$

Exercise 5.8.10

Find the line parallel to $y = -3x$ that passes through the point $(1, 1)$. Suggest another line and repeat.

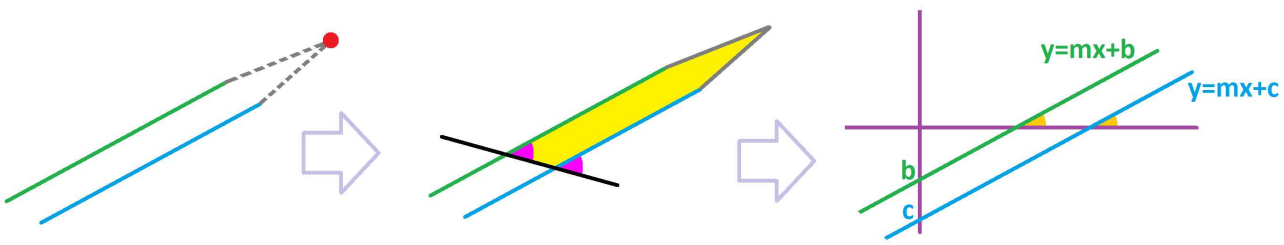
Exercise 5.8.11

Split the theorem into a statement and its converse.

Exercise 5.8.12

The case not covered by the theorem is a pair of two vertical lines. Show that they are all parallel to each other and not parallel to other lines.

Lines on the plane are its subsets. It is, therefore, natural to ask about their intersections. We know from Euclidean geometry that two parallel lines don't intersect. If they did, they'd form a triangle with the sum of the angles above 180 degrees:



The Cartesian system makes it possible to prove this fact algebraically.

Corollary 5.8.13: Parallel Lines: Basic Facts

- *Parallel lines don't intersect.*
- *Conversely, non-parallel lines intersect.*

Proof.

According to the last theorem, we have two lines with the same slope, m, and two different y-intercepts, b and c:

$$\begin{cases} y = mx + b & \texttt{AND} \\ y = mx + c\,. \end{cases}$$

For a point (x, y) to belong to the intersection, it would have to satisfy both. We have a system of linear equations! Subtracting the two equations produces: $b - c = 0$. There is no solution and, therefore, no intersection.

Exercise 5.8.14

Prove the converse part of the corollary.

Exercise 5.8.15

State the corollary as an equivalence (an "if-and-only-if" statement).

Exercise 5.8.16

Prove that vertical lines don't intersect each other and do intersect all other lines.

Example 5.8.17: mixtures, what can happen

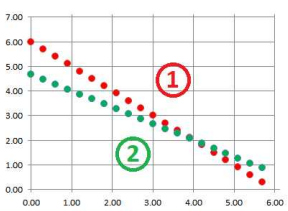
Recall an example from Chapter 2: Is it possible to create, from the Kenyan coffee ($2 per pound) and the Colombian coffee ($3 per pound), 6 pounds of blend with a total price of $14? The problem is solved via a system of linear equations:

$$\begin{cases} x & +y & = 6 \\ 2x & +3y & = 14\,. \end{cases} \quad \text{AND}$$

Without even solving it, we follow this line of thought:

- The slopes of the two lines are different; therefore,
- the lines are not parallel; therefore,
- the lines intersect; therefore,
- the system has a solution.

Confirmed:



So, it is *possible* to create such a blend! It would be impossible if both types of coffee were priced at $2 per pound.

Exercise 5.8.18

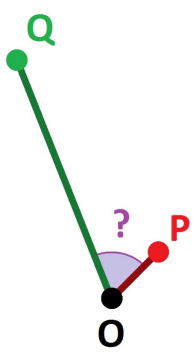
Justify the last statement – second possibility – using the theorem about parallel lines. What is the third possibility?

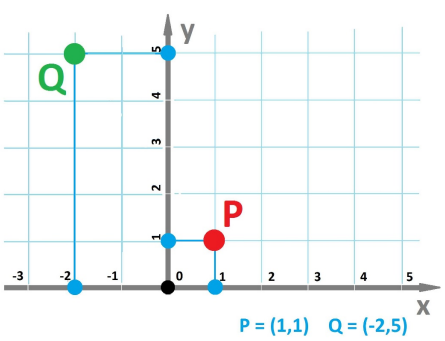
The value of the *angle between the directions* (i.e., the lines) from O to point P and from O to point Q comes from the triangle OPQ. This triangle and all of its measurements is inherited from the Euclidean plane that underlines the Cartesian system. The angle is denoted by $\widehat{QOP}$. Now, the question is:

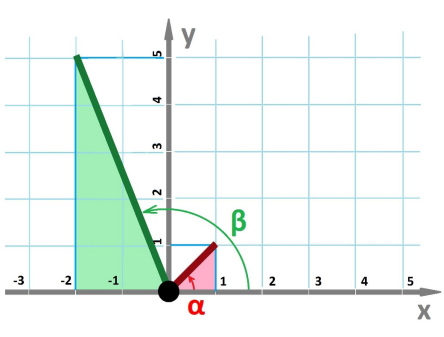
▶ How do we express $\widehat{QOP}$ in terms of their coordinates $P = (x, y)$ and $Q = (x', y')$?

Above we considered a special case: $Q = (1, 0)$.

The geometry is illustrated below:







$$\cos\alpha = \frac{1}{\sqrt{1^2+1^2}}$$

$$\cos\beta = \frac{-2}{\sqrt{2^2+5^2}}$$

We have the following from the theorem above. The two angles from the x-axis to $P = (x, y)$ and to

$Q = (x', y')$, respectively, satisfy the following:

$$\cos\alpha = \frac{x}{\sqrt{x^2+y^2}}\,, \quad \cos\beta = \frac{x'}{\sqrt{x'^2+y'^2}}$$
$$\sin\alpha = \frac{y}{\sqrt{x^2+y^2}}\,, \quad \sin\beta = \frac{y'}{\sqrt{x'^2+y'^2}}$$

The formulas look complicated but keep in mind that the denominators are just the distances from O to P and Q, respectively.

However, our interest isn't these two angles but their difference,

$$\widehat{QOP} = \alpha - \beta\,.$$

Fortunately, there is another trigonometric formula – Sine and Cosine of Difference – that allows us to represent the cosine of this angle in terms of the four quantities above:

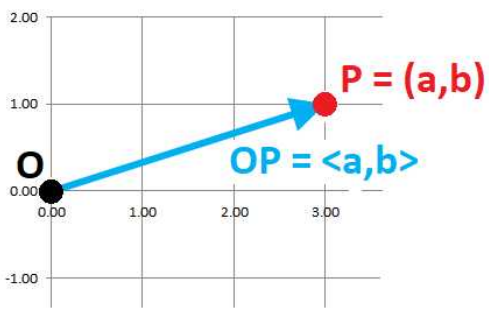
$$\begin{aligned}
\cos(\alpha-\beta) &= \cos\alpha\cos\beta + \sin\alpha\sin\beta \\
&= \frac{x}{\sqrt{x^2+y^2}}\frac{x'}{\sqrt{x'^2+y'^2}} + \frac{y}{\sqrt{x^2+y^2}}\frac{y'}{\sqrt{x'^2+y'^2}} && \text{We substitute.} \\
&= \frac{xx'+yy'}{\sqrt{x^2+y^2}\sqrt{x'^2+y'^2}}\,. && \text{And simplify.}
\end{aligned}$$

The two parts of the denominator are the distances to $P = (x, y)$ and $Q = (x', y')$:

$$d(O,P) = \sqrt{x^2+y^2} \quad \text{and} \quad d(O,Q) = \sqrt{x'^2+y'^2}\,.$$

But what about the numerator?

We take the vector approach again:



Definition 5.8.19: vector in dimension 2

If a segment's starting point is the origin, i.e., it's OP for some $P \neq O$, it is called a (2-dimensional) *vector* in $\mathbf{R}^2$. Its *components* are the coordinates of its terminal point P, according to the following notation:

$$P = (a,b) \iff OP =< a, b >$$

A vector has a direction, which is one of the two directions of the line, and a magnitude, defined as follows.

Definition 5.8.20: magnitude of vector

The *magnitude* of a vector $OP =< a, b >$ is defined as the distance from O to

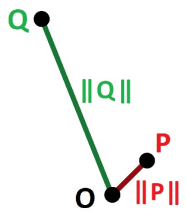
its tip P, denoted by

$$|| < a, b > || = \sqrt{a^2 + b^2}$$

Exercise 5.8.21

Apply the definition to a vector $< x, 0 >$ and explain its relation to the one-dimensional case.

The notation will help us with our formula for the angle:



It is rewritten as follows:

$$\cos(\alpha - \beta) = \frac{xx' + yy'}{|| < x, y > || \cdot || < x', y' > ||} .$$

We now would like to make sense of the numerator of this fraction. The following will become commonly used.

Definition 5.8.22: dot product

The *dot product* of two vectors $< a, b >$ and $< c, d >$ is defined as

$$< a, b > \cdot < c, d >= ac + bd$$

Warning!

Often, the meaning of the dot "$\cdot$" has to be determined from the context.

Exercise 5.8.23

Show that the dot product is linked back to the magnitude by the formula:

$$|| < a, b > ||^2 =< a, b > \cdot < a, b > .$$

The numerator of the formula for the angle $\cos(\alpha - \beta)$ takes a simpler form now:

$$< x, y > \cdot < x', y' > .$$

We conclude the following:

Theorem 5.8.24: Directions for Dimension 2

The angle between the vectors OP and OQ, where $P = (x, y) \neq O$ and $Q = (x', y') \neq O$, is determined by the following formula:

$$\cos \widehat{QOP} = \frac{< x, y > \cdot < x', y' >}{|| < x, y > || \, || < x', y' > ||}$$

So, the cosine of the angle can be now computed by using only addition, multiplication, and division of the four coordinates involved!

Exercise 5.8.25

Explain the difference between "the angle between two vectors" and "the angle between two lines".

Exercise 5.8.26

What is the angle between the lines $y = -2x + 3$ and $y = x - 1$? Suggest another pair of lines and repeat.

Exercise 5.8.27

Find a line that makes a 30-degree angle with the line $y = -2x + 3$? Suggest another angle and another line and repeat.

Exercise 5.8.28

Derive a formula for the sine of this angle.

Example 5.8.29: angle with itself

When two vectors are equal to each other, we have from the theorem:

$$\cos \widehat{POP} = \frac{< x, y > \cdot < x, y >}{|| < x, y > || \, || < x, y > ||} = \frac{xx + yy}{\sqrt{x^2 + y^2}\,\sqrt{x^2 + y^2}} = 1\,.$$

Therefore, $\widehat{POP} = 0$, as expected.

Example 5.8.30: angle within x-axis

When the terminal points of both vectors lie on the x-axis, i.e., $y = y' = 0$, the formula turns into the following:

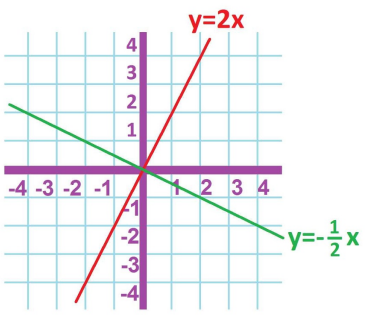
$$\cos \widehat{QOP} = \frac{xx'}{|x|\,|x'|} = \frac{x}{|x|}\frac{x'}{|x'|} = \text{sign}(x) \cdot \text{sign}(x')\,.$$

There are only two possibilities here, 1 or -1, and, therefore, $\widehat{QOP}$ can only be either 0 or 180 degrees.

Exercise 5.8.31

Show how this fact demonstrates the theorem about the directions in dimension 1.

A case special importance is: When are two vectors or two lines *perpendicular*? An example of these two lines, $y = 2x$ and $y = -\frac{1}{2}x$, suggests that the slopes will have to be *negative reciprocals* of each other:



Let's prove this fact using the theorem. We have

$$0 = \cos \pi/2 = \frac{< x, y > \cdot < x', y' >}{|| < x, y > || \, || < x', y' > ||}\,.$$

Therefore,

$$< x, y > \cdot < x', y' >= 0 \iff xx' + yy' = 0 \iff xx' = -yy' \iff \frac{y}{x} \cdot \frac{y'}{x'} = -1\,.$$

But these two expressions are the slopes of the lines!

Theorem 5.8.32: Slopes of Perpendicular Lines

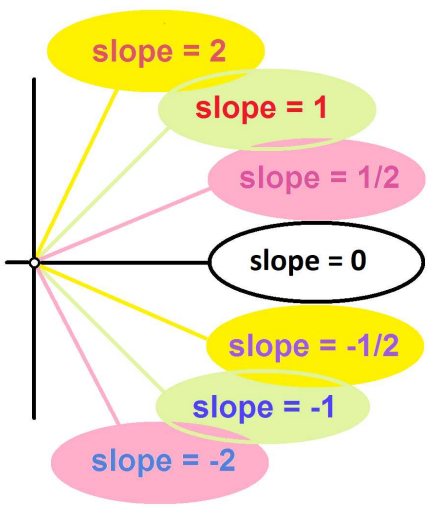
Two lines with slopes m and m' are perpendicular if and only if their slopes are negative reciprocals of each other; i.e.,

$$mm' = -1\,.$$

Exercise 5.8.33

Split the theorem into a statement and its converse.

Since any vertical line is perpendicular to any horizontal line and vice versa, we have solved the problem of perpendicularity. This is the summary:

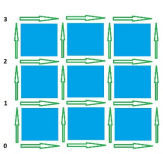


Exercise 5.8.34

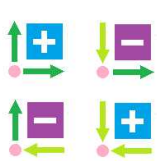
Find the line perpendicular to $y = -3x$ that passes through the point $(1, 1)$. Suggest another line and repeat.

The theory of vectors is further developed in Chapter 4HD-1.

The axes contain oriented segments. These segments then form *oriented rectangles* in the plane:



The ones above are all positively oriented, but, for example, combining a positively oriented segment and a negatively oriented segment will produce a negatively oriented rectangle:



So, since a piece of fabric has *two* sides, inside and outside, it can be placed on the table in *two* ways. This idea is further developed in Chapter 4HD-5.

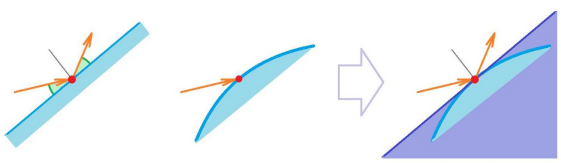
5.9. From geometry to calculus

There are two main entry points to calculus. The first one is via the study of *motion*. We followed this path in Chapter 1. The second path is via *geometry*. Starting in this section and then throughout calculus, we will

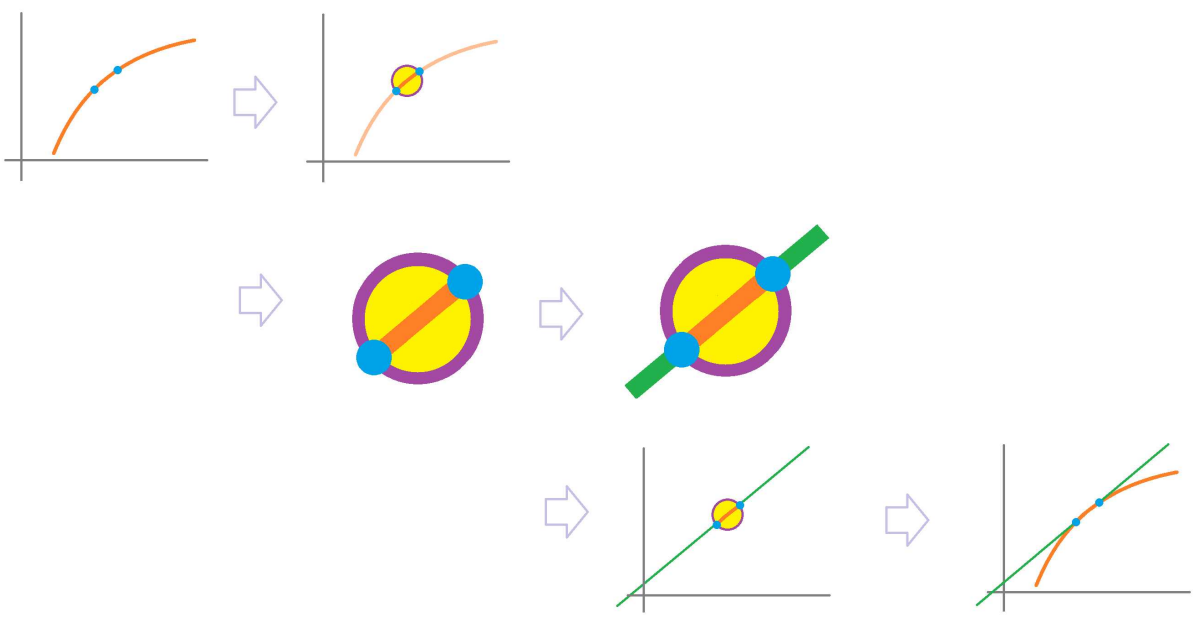
- compute *slopes* of curves using *differences*, and
- compute *areas* of curved regions using *sums*.

Let's have a preview.

First, in what direction will light bounce off a curved mirror? We can answer the question if we know the angle of the mirror at every location:



Let's zoom in on the point of contact. We might see a curve made of dots, similar to what we saw in Chapter 1. Every adjacent pair of points gives us a line, called the *secant line*:



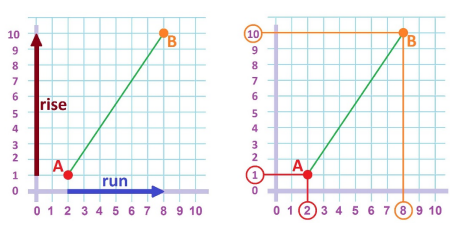
Then the *slope of the secant line* is seen to be, exactly or approximately, the slope of the curve.

Generally, suppose we have two points on the xy-plane:

$$(x_1, y_1) \quad \text{and} \quad (x_2, y_2)\,.$$

The *slope* of the line between them is known to be "rise over the run":

$$\text{slope} \;=\; \frac{y_2 - y_1}{x_2 - x_1}\,.$$



Of course, if we have more points, we have more slopes. Suppose a curve on the xy-plane is created from these sequences:

- x_n on the x-axis,
- y_n on the y-axis, and
- (x_n, y_n) on the plane.

The slopes of the curve are found as follows.

Recall the definition from Chapter 1: If x_n and y_n are two sequences, then their difference quotient is the difference of y_n over the difference of x_n; i.e.,

$$\frac{\Delta y_n}{\Delta x_n} = \frac{y_{n+1} - y_n}{x_{n+1} - x_n} .$$

Indeed, the numerator, the rise, is the change of y, i.e., the difference of y:

$$\Delta y_n = y_{n+1} - y_n .$$

The denominator, the run, is the change of x, i.e., the difference of x:

$$\Delta x_n = x_{n+1} - x_n .$$

Furthermore, recall the definition from Chapter 2: If $y = f(x)$ is a function and x_n is a sequence of points within its domain, then the *difference quotient* of f with respect to this sequence is defined to be the sequence of slopes of the lines from $(x_n, f(x_n))$ to $(x_{n+1}, f(x_{n+1}))$ on the graph of the function:

$$\boxed{\frac{\Delta f}{\Delta x} = \frac{f(x_{n+1}) - f(x_n)}{x_{n+1} - x_n}}$$

The sequence x_n is often chosen to be an arithmetic progression. In that case, its difference is the *increment* of the sequence:

$$h = \Delta x_n .$$

Then, the difference quotient is simply a multiple of the difference of $y_n = f(x_n)$:

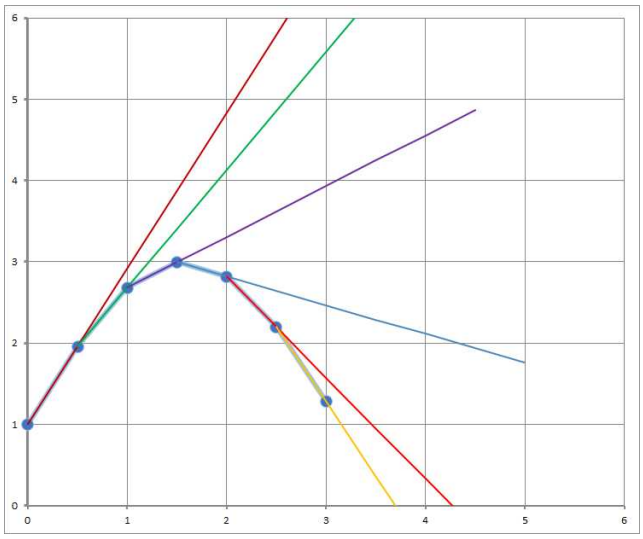
$$\frac{\Delta y_n}{\Delta x_n} = \frac{f(x_{n+1}) - f(x_n)}{h} = \frac{1}{h}\big(f(x_{n+1}) - f(x_n)\big) .$$

Example 5.9.1: slopes of parabola

Consider a parabola, say, $y = f(x) = -(x - 1.5)^2 + 3$. It is computed one value at a time with the following spreadsheet formula:

```
=(-1)*(RC[-1]-1.5)^2+3
```

and then plotted point by point:

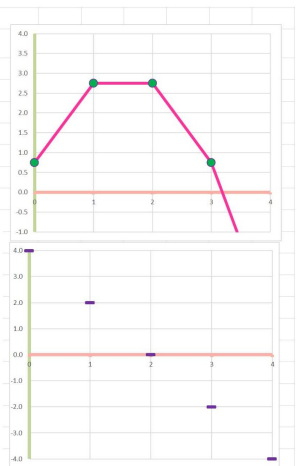


For each two consecutive pair of points, the slope is found by the formula, just as in Chapter 1:

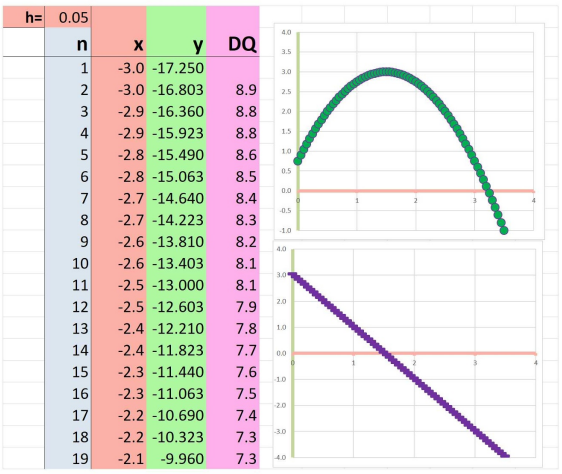
```
=(RC[-2]-R[-1]C[-2])/(RC[-1]-R[-1]C[-1])
```

The line is then drawn via the *point-slope formula*. In search of a pattern, we make the points denser and denser:

h=	1.00			
	n	x	y	DQ
	1	-3.0	-17.250	
	2	-2.0	-9.250	8.0
	3	-1.0	-3.250	6.0
	4	0.0	0.750	4.0
	5	1.0	2.750	2.0
	6	2.0	2.750	0.0
	7	3.0	0.750	-2.0



h=	0.05			
	n	x	y	DQ
	1	-3.0	-17.250	
	2	-3.0	-16.803	8.9
	3	-2.9	-16.360	8.8
	4	-2.9	-15.923	8.8
	5	-2.8	-15.490	8.6
	6	-2.8	-15.063	8.5
	7	-2.7	-14.640	8.4
	8	-2.7	-14.223	8.3
	9	-2.6	-13.810	8.2
	10	-2.6	-13.403	8.1
	11	-2.5	-13.000	8.1
	12	-2.5	-12.603	7.9
	13	-2.4	-12.210	7.8
	14	-2.4	-11.823	7.7
	15	-2.3	-11.440	7.6
	16	-2.3	-11.063	7.5
	17	-2.2	-10.690	7.4
	18	-2.2	-10.323	7.3
	19	-2.1	-9.960	7.3



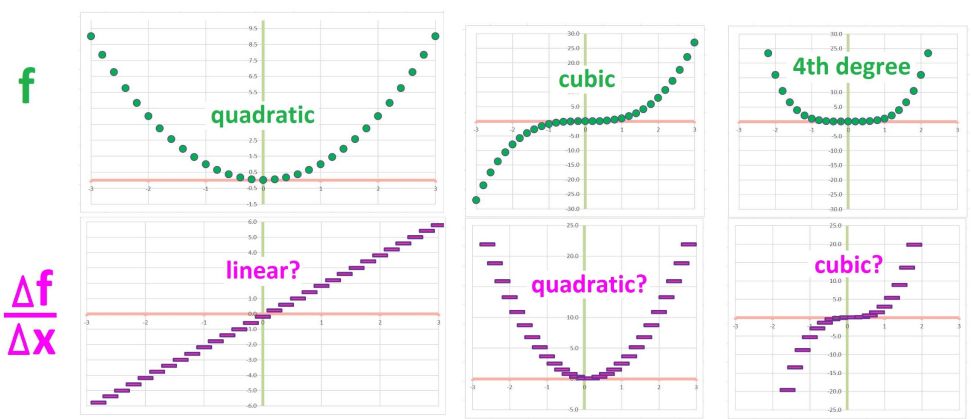
As we produce more and more points on this interval, we realize that the difference quotient is a straight line!

Exercise 5.9.2

Prove the last statement.

We have followed this pattern several times in the past.

We will see more patterns if we collect the difference quotients of the functions we have seen previously. These are those of the power functions:

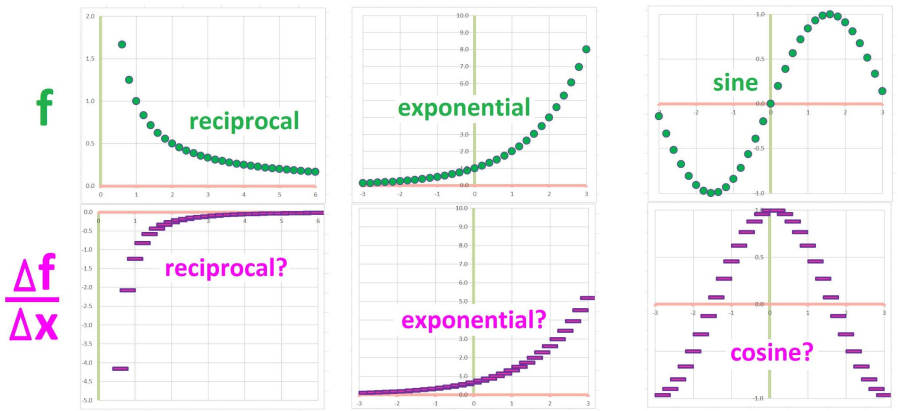


It appears that the power goes down by one!

Exercise 5.9.3

Sketch the next pair.

The rest of the functions also seem to be connected to each other:



Exercise 5.9.4

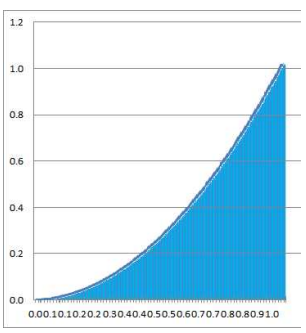
Try to answer these questions.

Exercise 5.9.5

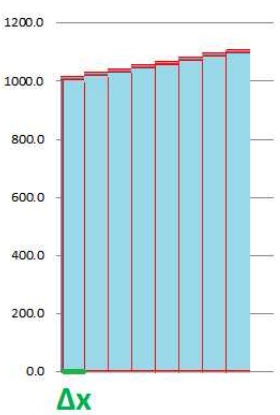
Use the formula above to find and plot the difference quotient for the circle.

The patterns suggested by this data and these pictures are studied by calculus. But, furthermore, how do we treat motion when the time varies *continuously* instead of incrementally? What is the meaning of the velocity then? Calculus also gives the answer (Chapter 2DC-3).

Second, how do we find the *area* of a curved region?



When we zoom in, we might see that the curve is made of a sequence of points, just as above. The curve on the xy-plane is created by two sequences, (x_n, y_n). At its simplest, the region is made of *rectangles*:



The area of each of them is the width times the height, which is the value of the function at that x:

$$\Delta x_n \cdot y_n = \Delta x_n \cdot f(x_n)\,.$$

Now, the total area is the sum of the areas of the rectangles; we use the *sums of sequences* just as in Chapter 1:

$$A = \sum_{n=1}^{m} \Delta x_n \cdot f(x_n) = \sum_{n=1}^{m} f(x_n)\,\Delta x_n$$

It is called the *Riemann sum*.

The sequence x_n is often chosen to be an arithmetic progression. In that case, its difference is the *increment* of the sequence:

$$h = \Delta x_n\,.$$

Then, the Riemann sum is simply a multiple of the sum of the sequence $y_n = f(x_n)$:

$$A = \sum_{n=1}^{m} f(x_n)\,\Delta x_n = h \sum_{n=1}^{m} f(x_n)\,.$$

Example 5.9.6: area of circle

We know that the area of a circle of radius r is supposed to be $A = \pi r^2$. Let's try to confirm this with nothing but a spreadsheet. First, we plot the graph

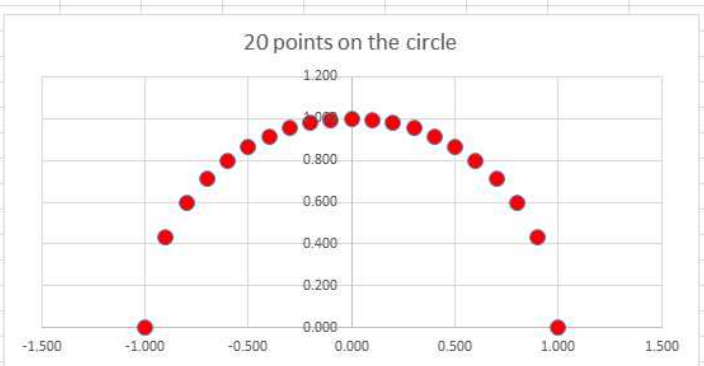
$$y = \sqrt{1 - x^2}\,,$$

by letting the values of x run from -1 to 1 every .1 and finding the values of y with the spreadsheet formula:

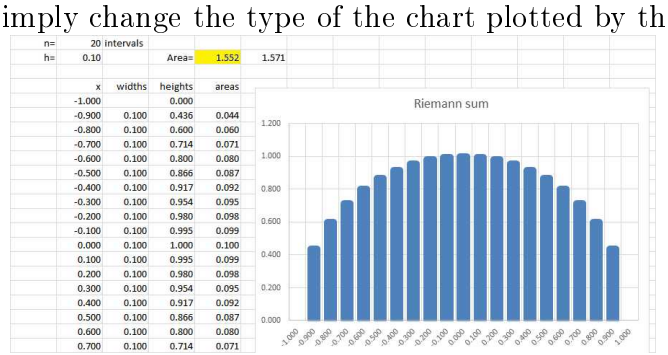
```
=SQRT(1-RC[-2]^2)
```

We plot these 20 points; the result is a half-circle:

x	y
-1.000	0.000
-0.900	0.436
-0.800	0.600
-0.700	0.714
-0.600	0.800
-0.500	0.866
-0.400	0.917
-0.300	0.954
-0.200	0.980
-0.100	0.995
0.000	1.000
0.100	0.995
0.200	0.980
0.300	0.954
0.400	0.917
0.500	0.866
0.600	0.800
0.700	0.714
0.800	0.600



We next cover, as best we can, this half-circle with vertical bars that stand on the interval $[-1, 1]$. We re-use the data: The bases of the bars are our intervals in the x-axis, and the heights are the values of y. To see the bars, we simply change the type of the chart plotted by the spreadsheet:



Then, the area of the circle is approximated by the *sum* of the areas of the bars! To let the spreadsheet do the work for us: Multiply the heights by the (constant) widths in the next column and add them up. We use the formulas for *sums* just as in Chapter 1:

```
=R[-1]C+RC[-1]*(RC[-2]-R[-1]C[-2])
```

The result produced is the following:

$$\text{Approximate area of the semicircle} = 1.552\,.$$

It is close to what we know. In summary, the area A is the sum of the sequence:

$$a_n = \sqrt{1-t_n^2}\cdot .1\,, \quad n = 1, 2, ..., 20\,.$$

In other words:

$$A = \sum_{n=1}^{20} .1\cdot\sqrt{1-t_n^2} = \left(\sum_{n=1}^{20}\sqrt{1-t_n^2}\right)\cdot .1\,.$$

The theoretical result will be proven in Chapter 3IC-1.

Exercise 5.9.7

Use the formula above to find the area under the parabola in the first example.

5.10. Solving inequalities

As explained in Chapter 2, to solve an inequality with respect to x means the same as solving an equation – to find *all* values of x that, when substituted, produce a true statement. These numbers form a set called the solution set.

There is a difference, of course; in the case of an equation, such a statement is likely to be

$$0 = 0,\ 0 = 1,\ 100 = 100,\ \text{etc.}$$

or similar, while in the case of an inequality, there might be a variety of these:

$$0 < 1,\ 0 > 1,\ 0 < 100,\ 100 \le 100,\ \text{etc.}$$

Consider how this inequality is solved:

$$x + 2 \ge 5 \implies x \ge 3\,.$$

That's an abbreviated version of the following statement:

- If x satisfies the inequality $x + 2 \ge 5$, then x satisfies the inequality $x \ge 3$.

Plug in some values for x and see if they check out:

	$(x)+2=$	$\ge^?\ 5$	TRUE/FALSE	
$x=0$	$(0)+2=2$	$\not\ge 5$	FALSE	
$x=1$	$(1)+2=3$	$\not\ge 5$	FALSE	
$x=2$	$(2)+2=4$	$\not\ge 5$	FALSE	
$x=3$	$(3)+2=5$	≥ 5	TRUE	Add it to the list!
$x=4$	$(4)+2=6$	≥ 5	TRUE	Add it to the list!

Of course, this trial-and-error method is unfeasible because there are infinitely many possibilities.

Recall the basic methods, i.e., rules, of *handling* inequalities.

Theorem 5.10.1: Basic Algebra of Inequalities

- *Multiplying both sides of an inequality by a positive number preserves it:*

$$a < b \implies ka < kb \quad \text{for any } k > 0\,.$$

- *Multiplying both sides of an inequality by a negative number reverses it:*

$$a < b \implies ka > kb \quad \text{for any } k < 0\,.$$

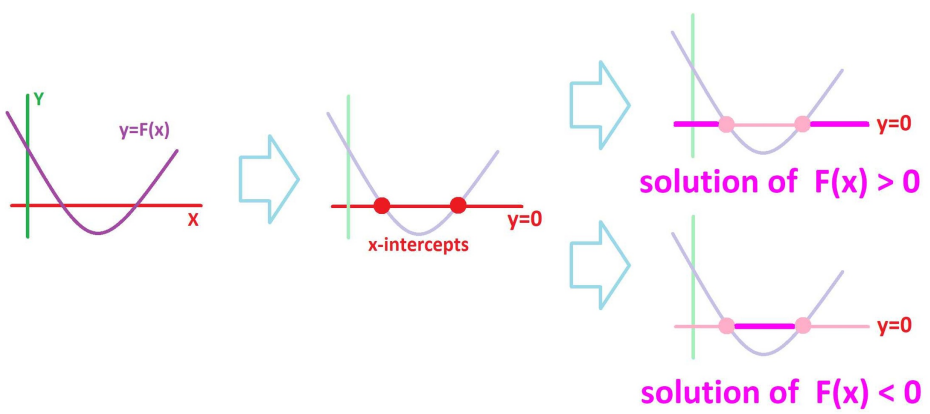
- *Adding any number to both sides of an inequality preserves it:*

$$a < b \implies a + s < b + s \quad \text{for any } s\,.$$

Warning!

Multiplying an inequality by 0 is pointless.

As a reminder, in the special case when the right-hand side of the inequality is zero, the solution set to this inequality has a clear *geometric* meaning: It is comprised of the intervals cut by the *x-intercepts*:

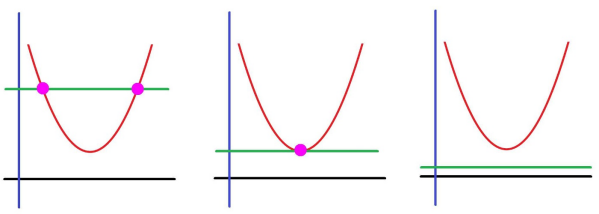


Exercise 5.10.2

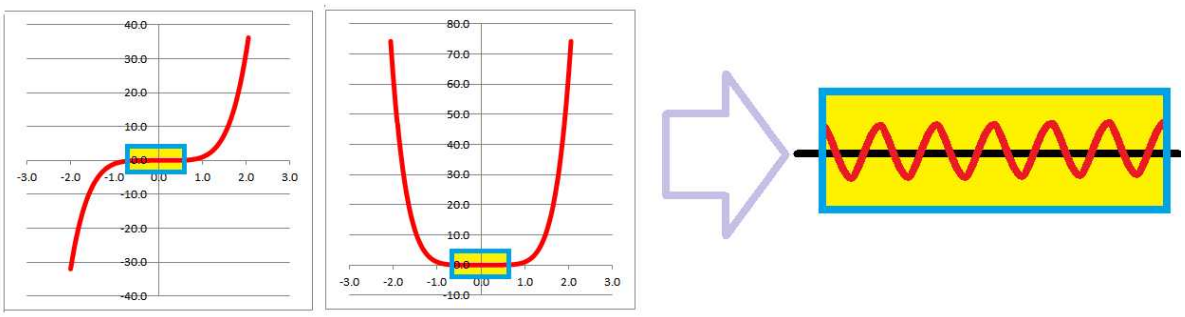
These solution sets are the intersections of _____ with the x-axis.

Example 5.10.3: counting intervals

Our knowledge of the shape of the graph of a particular function will tell us how many solutions such an inequality might have. For example, a quadratic inequality can have two, one, or no intervals:



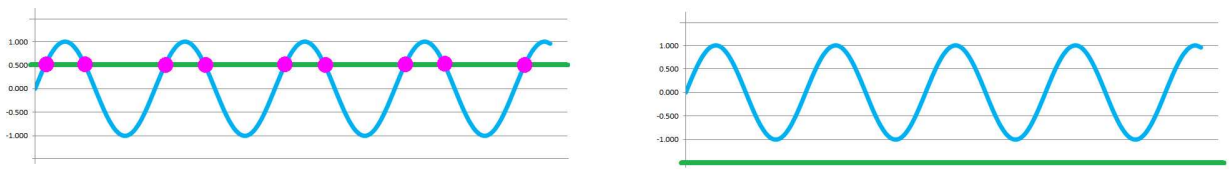
An equation with an nth degree polynomial cannot have more than $n + 1$ inequalities:



An inequality with the exponential function can have one interval or none:



An inequality with a periodic function (such as the sine) can have infinitely many intervals or none:



Exercise 5.10.4

Sketch the solution set for the inequalities $f(x) > 0$ and $f(x) < 0$ with the functions f given in the last example.

A method of *how* we may arrive to the answer is discussed in this section. The analysis follows the one for *solving equations* presented earlier in this chapter.

We will address a simple kind of inequality:

- x is present only once (in the left-hand side).

Like this:

$$x^2 \geq 17\,.$$

Starting with such an equation, our goal is – through a series of manipulations – to arrive to an even simpler kind of inequality:

- x is *isolated* (in the left-hand side).

Like this:

$$x \geq \sqrt{17}\,.$$

What we see above is a single x "wrapped" in several layers of functions, i.e., a composition of several functions applied to x. We will remove these layers one by one, from the outside in.

Warning!

This plan does not work for many familiar types of inequalities considered in Chapter 4.

The main idea is the same as for equations:

▶ We apply a function to both sides of the inequality, producing a new inequality with a possibly reversed sign.

For example, if we have an inequality, say,

$$x+2 \geq 5\,,$$

we treat the two sides of the inequalities as two values of the same variable, say, y. If we add 2 to both, the relation between them remains. More precisely, we have the following:

$$x+2 \geq 5, \quad \text{apply } z = y-2 \implies (x+2)-2 \geq 3-2 \implies x \geq 3\,.$$

From an inequality satisfied by x, we produce another inequality satisfied by x. The solution set is $[3, +\infty)$

However, the challenge is that *any* function applied this way may produce a new inequality! More precisely, any *increasing* function will automatically produce a new, correct, inequality:

$$\begin{array}{lll} x+2 \geq 5, \quad \text{apply } z = y+2 & \implies (x+2)+2 \geq 5+2 \implies x+4 \geq 7\,. & \\ x+2 \geq 5, \quad \text{apply } z = 7y & \implies 7\cdot(x+2) \geq 7\cdot 5\,. & \text{Not solved!} \\ x+2 \geq 5, \quad \text{apply } z = y^3 & \implies (x+2)^3 \geq 5^3\,. & \text{Not solved!} \end{array}$$

This is the *definition* of an increasing function: Larger inputs (an inequality) produce, under g, larger outputs (another inequality). That's what we have on the left, and the opposite on the right:

$$\begin{array}{lccc} \text{old inequality:} & a & \geq & b \\ \boxed{g\nearrow} & \downarrow g & & \downarrow g \\ \text{new inequality:} & g(a) & \geq & g(b) \end{array} \qquad \begin{array}{lccc} \text{old inequality:} & a & \geq & b \\ \boxed{g\searrow} & \downarrow g & & \downarrow g \\ \text{new inequality:} & g(a) & \leq & g(b) \end{array}$$

So, we have, for any increasing function g:

$$x+2 \geq 5 \implies g(x+2) \geq g(5)\,.$$

The opposite is true for any decreasing function g:

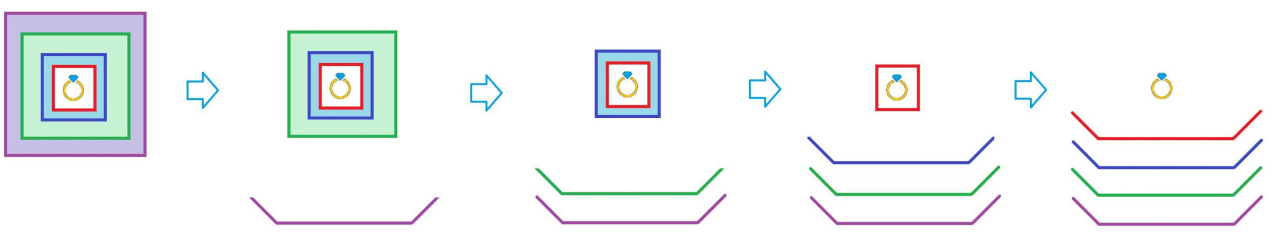
$$x+2 \geq 5 \implies g(x+2) \leq g(5)\,.$$

Unfortunately, there are infinitely many possibilities for this function g:

$$\begin{array}{ccccc} x+4 \geq 7 & & 7\cdot(x+2) \geq 7\cdot 5 & & (-2)\cdot(x+2) \leq (-2)\cdot 5 \\ & \nwarrow & \uparrow & \nearrow & \\ x+5 \geq 8 & \leftarrow & \boxed{x+2 \geq 5} & \rightarrow & 2^{x+2} \geq 2^5 \\ & \swarrow & \downarrow & \searrow & \\ x+6 \geq 9 & & (x+2)^3 \geq 5^3 & & \sqrt{x+2} \geq \sqrt{5} \end{array}$$

From these, the only decreasing function is $(-2)y$.

If x satisfies the inequality in the middle, it also satisfies the rest of the inequalities. If we want to solve the original inequality, we will need a foresight to choose a function to apply that will make the inequality *simpler*. Just as with equations, if between us and x there is a function, we remove it by applying its *inverse* to the inequality. We work from outside in (while the gift-giver worked from the inside out):



Example 5.10.5: solving inequality

What if we have *several* functions applied consecutively to x? Which function do we choose to apply? In the inequality,

$$5 \cdot \left(\left(\frac{x}{2} + 1 \right)^2 + 3 \right) - 17 \geq 3 \,,$$

the last operation on the left is -17. That's the function we face, and the inverse of this function is to be applied. We choose, therefore:

$$f(z) = z - 17 \,,$$

where

$$z = 5 \cdot \left(\left(\frac{x}{2} + 1 \right)^2 + 3 \right) .$$

Then we apply

$$g(y) = f^{-1}(y) = y + 17 \,.$$

to both sides. We conclude:

$$z - 17 \geq 3 \implies (z - 17) + 17 \geq 3 + 17 \implies z \geq 20 \,.$$

We have a new inequality now:

$$5 \cdot \left(\left(\frac{x}{2} + 1 \right)^2 + 3 \right) \geq 20 \,.$$

As we progress, we apply the inverse of the function that appears first in the right-hand side of the inequality. We have the following sequence of steps removing functions one at a time:

$$\begin{aligned} &5 \cdot \left(\left(\frac{x}{2} + 1 \right)^2 + 3 \right) \geq 20 \implies 5 \cdot \left(\left(\frac{x}{2} + 1 \right)^2 + 3 \right) / 5 \geq 20/5 \implies \\ &\left(\frac{x}{2} + 1 \right)^2 + 3 \geq 4 \implies \left(\frac{x}{2} + 1 \right)^2 + 3 - 3 \geq 4 - 3 \implies \\ &\left(\frac{x}{2} + 1 \right)^2 \geq 1 \,. \end{aligned}$$

The last inequality produces two cases depending on which branch of $y = x^2$ we consider:

$$\left(\frac{x}{2} + 1 \right)^2 \geq 1$$

$\swarrow$ $\qquad$ $\searrow$

$z = \frac{x}{2} + 1 \geq 0$			OR	$z = \frac{x}{2} + 1 \leq 0$		
$\left(\frac{x}{2} + 1 \right)$	≥ 1	$\implies$	OR	$-\left(\frac{x}{2} + 1 \right)$	≥ 1	$\implies$
$\frac{x}{2} + 1$	≥ 1	$\implies$	OR	$\frac{x}{2} + 1$	≤ -1	$\implies$
$\frac{x}{2}$	≥ 0	$\implies$	OR	$\frac{x}{2}$	≤ -2	$\implies$
x	≥ 0		OR	x	≤ -4	

The values of x, and only they, that belong to $[0, +\infty)$ or to $(-\infty, -4]$ satisfy this restriction. Therefore, the solution set is the *union* of these two intervals:

$$(-\infty, -4] \cup [0, +\infty) \,.$$

Exercise 5.10.6

Solve the inequality:

$$5 \cdot \left(\left(\frac{x^2}{2} + 1 \right)^2 + 3 \right) - 17 \geq 3 \, .$$

Exercise 5.10.7

Solve the inequality:

$$\left(5 \cdot \left(\left(\frac{x}{2} + 1 \right)^2 + 3 \right) - 17 \right)^3 \geq 27 \, .$$

Make up your own inequality and solve it. Repeat.

Using invertible functions ensures that we will have *the same solution set* as we progress through the stages. For example, this is how we would rather present the solution of the very first inequality in this section:

$$x + 2 \geq 5 \iff x \geq 3 \, .$$

That's an abbreviated version of the following statement:

$$x \text{ satisfies } x + 2 \geq 5 \text{ if and only if } x \text{ satisfies } x \geq 3.$$

Repeated as many times as necessary, the solution set of each inequality is the same as that of the original inequality!

Example 5.10.8: split domain

Two complete solutions are below:

$$\begin{array}{ll} (1) & 2(x+2) - 3 \geq 5x \iff 2x + 4 - 3 \geq 5x \iff -3x \geq -1 \iff x \leq 1/3 \, . \\ (2) & x^2 + 1 \geq 0 \iff x^2 \geq -1 \iff \mathbf{R} \, . \end{array}$$

So, to solve the type of inequality we face – a variable x subjected to a sequence of functions – we apply the inverses of these functions in the reversed order. We may have to split the domain of the function so that it is one-to-one on each of the subsets. As we have seen, this step also splits our inequality.

As a summary, this is our method.

Theorem 5.10.9: General Algebra of Inequalities

- *If g is a strictly increasing function, then applying g to both sides of an inequality creates an equivalent inequality:*

$$a < b \iff g(a) < g(a) \, .$$

- *If g is an strictly decreasing function, then applying g to both sides of an inequality creates an equivalent (but reversed) inequality:*

$$a < b \iff g(a) > g(a) \, .$$

In either case, the two inequalities have the same solution set.

Example 5.10.10: split domain

A solution is below:

$$x^2 \geq 1 \iff x \leq -1 \quad \texttt{OR} \quad x \geq 1 \, .$$

Exercise 5.10.11

Solve the inequality in this manner:

$$2^{x+1} \geq 3\,.$$

Exercise 5.10.12

Solve the inequality in this manner:

$$2^{x^2+1} \geq 3\,.$$

Make up your own inequality and solve it. Repeat.

Example 5.10.13: split domain

If we, instead of a single x, face an expression that depends on x, we treat it as just another variable:

$$5 \cdot \Big(\big(x^2+x+1\big)^2 + 3 \Big) - 17 \geq 3 \implies \text{choose } \; y = x^2+x+1\,.$$

We substitute:

$$5 \cdot \Big((y)^2 + 3 \Big) - 17 \geq 3\,,$$

and go after y by following the same plan as before. Once this inequality is solved, you have a much simpler inequality, or inequalities, for x:

$$x^2+x+1 \geq 1 \quad \texttt{OR} \quad x^2+x+1 \leq -4\,.$$

Exercise 5.10.14

Finish the solution. Make up your own inequality and solve it. Repeat.

Chapter : Exercises

Contents

1 Exercises: Algebra 487
2 Exercises: Sequences 490
3 Exercises: Sets and logic 491
4 Exercises: Coordinate system 492
5 Exercises: Linear algebra 493
6 Exercises: Polynomials 495
7 Exercises: Relations and functions 497
8 Exercises: Graphs 500
9 Exercises: Compositions 502
10 Exercises: Transformations 505
11 Exercises: Basic models 507

1. Exercises: Algebra

Exercise 1.1

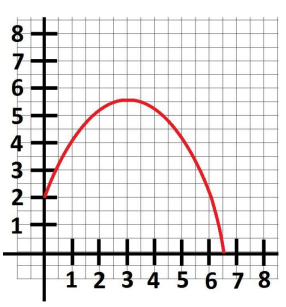
Solve the following equation: $\sqrt{x^2-7}-3=0$.

Exercise 1.2

Solve the following equation: $\left(x^2-7\right)^3-8=0$.

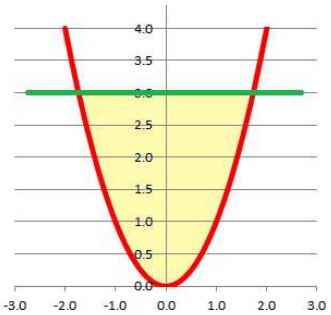
Exercise 1.3

Consider the parabola below. Find its vertex, its axis of symmetry, and its maximum or minimum.



Exercise 1.4

Find the exact values of the x-coordinates of the intersections between the parabola and the line shown below:



Exercise 1.5

Solve the equation: $2^x = 3^{x+1}$. Don't simplify.

Exercise 1.6

Solve the equation: $3^x = 3^{x+1}$.

Exercise 1.7

Solve the equation: $3^x = 2$.

Exercise 1.8

Solve the equation: $2^x = 2 \cdot 3^{x+1}$. Don't simplify.

Exercise 1.9

To what power should you raise 3 to get 10?

Exercise 1.10

Find the domain, the range, and the asymptotes of the function $f(x) = \ln(x-3) + \ln 3$.

Exercise 1.11

There are 125 sheep and 5 dogs in a flock. How old is the shepherd?

Exercise 1.12

Let $h(x) = x^2 + 3x - 10$. Find the x- and y-intercepts and sketch the graph of the function.

Exercise 1.13

Suggest an equation the solutions of which are 1 and 2.

Exercise 1.14

Expand:

$$(x+1)^5 = \ldots$$

Exercise 1.15

Simplify:

$$\left(2^{20} \cdot 2^{35}\right)^{10} = \ldots$$

Exercise 1.16

Solve these equations:

$$x^2 = x, \quad x = -x, \quad x = 0 \cdot x.$$

Exercise 1.17

Represent as a power of 5:

$$5^3 \cdot .2, \ (125 \cdot 5)^5.$$

Exercise 1.18

If bacteria triple in number every day and the current population is 9000, how many were there three days ago?

Exercise 1.19

Factor $2a^2 + 12ab + 18b^2$.

Exercise 1.20

Contract this summation:

$$2 - \frac{2}{2} + \frac{2}{3} - \frac{2}{4} = ?$$

Exercise 1.21

Expand this summation:

$$\sum_{k=-1}^{5} \frac{k^2}{k+2} = ?$$

Exercise 1.22

True or False?
(a) $x < r \implies |x| < r$;
(b) $x < r \impliedby |x| < r$;
(c) $x < r \iff |x| < r$.

Exercise 1.23

For $4y + 16x = 20$, provide the slope and the y-intercept.

Exercise 1.24

Let $f(x) = 4x^2 + 2x + 2$ and let

$$g(h) = \frac{f(1+h) - f(1)}{h}.$$

Determine each of the following:

1. $g(1) = \ldots$
2. $g(0.1) = \ldots$
3. $g(0.01) = \ldots$

What is the trend?

Exercise 1.25

A diameter of a circle runs between points R and T. The center of the circle, P, has coordinates $(-4, 1)$. The coordinates of the point R are $(2, -3)$. What are the coordinates of T?

Exercise 1.26

Let $f(x) = 7+2x-x^2$. Find the difference quotient

$$\frac{f(3+h) - f(3)}{h}.$$

Simplify your answer.

Exercise 1.27

Find an expression for $f(x)$ and state its domain in interval notation given that f is the function that takes a real number x and performs the following three steps in order:

1. divide by 3,
2. take square root, and then
3. make the quantity the denominator of a fraction with numerator 13.

Exercise 1.28

Solve the equation:

$$\sin x = \cos x.$$

2. Exercises: Sequences

Exercise 2.1

Compute $\sum_{n=1}^{4} n^2$.

Exercise 2.2

Present the first 5 terms of the sequence:

$$a_1 = 1, \quad a_{n+1} = -(a_n + 1).$$

Exercise 2.3

Represent in sigma notation:

$$-1 - 2 - 3 - 4 - 5 - ... - 10.$$

Exercise 2.4

Find the sum of the following:

$$-1 - 2 - 3 - 4 - 5 - ... - 10.$$

Exercise 2.5

Find the sequence of sums of the following sequence:

$$-1, 2, -4, 8, -5, ...$$

Exercise 2.6

Show that $\frac{n}{n+1}$ is an increasing sequence. What kind of sequence is $\frac{n+1}{n}$? Give examples of increasing and decreasing sequences.

Exercise 2.7

Find the next item in each list:

1. $7, 14, 28, 56, 112, ...$
2. $15, 27, 39, 51, 63, ...$
3. $197, 181, 165, 149, 133, ...$

Exercise 2.8

A pile of logs has 50 logs in the bottom layer, 49 logs in the next layer, 48 logs in the next layer, and so on, until the top layer has 1 log. How many logs are in the pile?

Exercise 2.9

In the beginning of each year, a person puts \$5000 in a bank that pays 3% compounded annually. How much does he have after 15 years?

Exercise 2.10

An object falling from rest in a vacuum falls approximately 16 feet the first second, 48 feet the second second, 80 feet the third second, 112 feet the fourth second, and so on. How far will it fall in 11 seconds?

3. Exercises: Sets and logic

Exercise 3.1

Represent the following set in the set-building notation:

$$X = [0, 1] \cup [2, 3] = ...$$

Exercise 3.2

Simplify:

$$\{x > 0 : \ x \text{ is a negative integer } \}.$$

Exercise 3.3

What are the max, min, and any bounds of the set of integers? What about $\mathbf{R}$?

Exercise 3.4

Is the converse of the converse of a true statement true?

Exercise 3.5

State the converse of this statement: "the converse of the converse of a true statement is true".

Exercise 3.6

Represent these sets as intersections and unions:

1. $(0, 5)$
2. $\{3\}$
3. $\emptyset$
4. $\{x : x > 0 \quad \texttt{OR} \quad x \text{ is an integer}\}$
5. $\{x : \ x \text{ is divisible by } 6\}$

Exercise 3.7

True or false: $0 = 1 \implies 0 = 1$?

Exercise 3.8

Prove:

$$\max\{\max A, \max B\} = \max(A \cup B).$$

Exercise 3.9

(a) If, starting with a statement A, after a series of conclusions you arrive to $0 = 1$, what can you conclude about A? (b) If, starting with a statement A, after a series of conclusions you arrive to $0 = 0$, what can you conclude about A?

Exercise 3.10

We know that "If it rains, the road gets wet". Does it mean that if the road is wet, it has rained?

Exercise 3.11

A garage light is controlled by a switch and, also, it may automatically turn on when it senses motion during nighttime. If the light is OFF, what do you conclude?

Exercise 3.12

If an advertisement claims that "All our second-hand cars come with working AC", what is the easiest way to disprove the sentence?

Exercise 3.13

Teachers often say to the student's parents: "If your student works harder, he'll improve". When he won't improve and the parents come back to the teacher, he will answer: "He didn't improve, that means he didn't work harder". Analyze.

4. Exercises: Coordinate system

Exercise 4.1

Find the equation of the line passing through the points $(-1, 1)$ and $(-1, 5)$.

Exercise 4.2

What is the distance from the center of the circle

$$(x-1)^2 + (y+3)^2 = 5$$

to the origin?

Exercise 4.3

What is the distance from the circle

$$x^2 + (y+3)^2 = 2$$

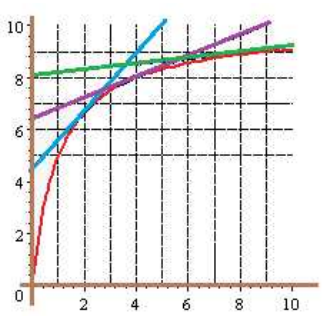
to the origin?

Exercise 4.4

Find the equation of the circle centered at $(-1, -1)$ and passing through the point $(-1, 1)$.

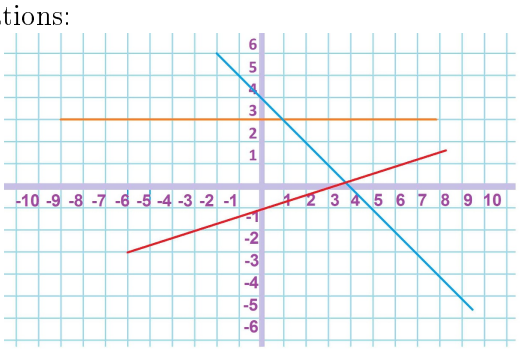
Exercise 4.5

Three straight lines are shown below. Find their slopes:



Exercise 4.6

Three straight lines are shown below. Find their equations:



Exercise 4.7

For the points $P = (0, 1)$, $Q = (1, 2)$, and $R = (-1, 2)$, determine the points that are symmetric with respect to the axis and the origin.

Exercise 4.8

The hypotenuse of an isosceles right triangle is 10 inches. The midpoints of its sides are connected to form an inscribed triangle, and this process is repeated. Find the sum of the areas of these triangles as this process is continued.

Exercise 4.9

Consider triangle ABC in the plane where $A = (3, 2)$, $B = (3, -3)$, $C = (-2, -2)$. Find the lengths of the sides of the triangle.

Exercise 4.10

Sketch the region given by the set $\{(x, y) : xy < 0\}$. Which axes and which quadrants of the plane are included in the set?

Exercise 4.11

Find all x such that the distance between the points $(3, -8)$ and $(x, -6)$ is 5.

Exercise 4.12

Two care leave a highway junction at the same time. The first travels west at 70 miles per hour and the second travels north at 60 miles per hour. How far apart are they after 1.5 hours?

Exercise 4.13

Find the perimeter of the triangle with the vertices at $(3, -1)$, $(3, 6)$, and $(-6, -5)$.

Exercise 4.14

Find the point on the x-axis that is equidistant from the points $(-1, 5)$ and $(6, 4)$.

Exercise 4.15

(a) Give the definition of the circle of radius R centered at a point C. (b) Find the equation of the circle centered at the point $(1, 1)$ that passes through $(0, 0)$.

5. Exercises: Linear algebra

Exercise 5.1

What is the equation of the line through the points $A = (-3, 2)$ and $B = (2, 5)$?

Exercise 5.2

Find the distance between the points of intersection of the circle $(x-1)^2 + (y-2)^2 = 6$ with the axes.

Exercise 5.3

Set up, but do not solve, a system of linear equations for the following problem: "Suppose your portfolio is worth $\$20,000$ and it consists of two stocks A and B. The stocks are priced as follows: A \$2.1 per share, B \$1.5 per share. Suppose also that you have twice as much of stock A than B. How much of each do you have?"

Exercise 5.4

In an effort to find the point in which the lines $2x - y = 2$ and $-4x + 2y = 1$ intersect, a student multiplied the first one by 2 and then added the result to the second. He got $0 = 5$. Explain the result.

Exercise 5.5

Find the angle between the lines: from $(0, 0)$ to $(1, 1)$ and from $(0, 0)$ to $(1, 2)$. Don't simplify.

Exercise 5.6

Solve the system of linear equations:

$$\begin{cases} x & -y & = -1, \\ 2x & +y & = 0. \end{cases}$$

Exercise 5.7

Solve the system of linear equations:

$$\begin{cases} x & -2y & = 1, \\ 2x & +y & = 0. \end{cases}$$

Exercise 5.8

A movie theater charges \$10 for adults and \$6 for children. On a particular day when 320 people paid an admission, the total receipts were \$3120. How many were adults and how many were children?

Exercise 5.9

The taxi charges \$1.75 for the first quarter of a mile and \$0.35 for each additional fifth of a mile. Find a linear function which models the taxi fare f as a function of the number of miles driven, x.

Exercise 5.10

Given vectors $a =< 1, 2 >$, $b =< -2, 1 >$, find their magnitudes and the angle between them.

Exercise 5.11

Set up a system of linear equations – but do not solve – for the following problem: "A mix of coffee is to be prepared from: Kenyan coffee - \$3 per pound and Colombian coffee - \$5 per pound. How much of each do you need to have 10 pounds of blend with \$3.50 per pound?"

Exercise 5.12

Set up, do not solve, the system of linear equations for the following problem: "One serving of tomato soup contains 100 Cal and 18 g of carbohydrates. One slice of whole bread contains 70 Cal and 13 g of carbohydrates. How many servings of each should be required to obtain 230 Cal and 42 g of carbohydrates?"

Exercise 5.13

Solve the system of linear equations:

$$\begin{cases} x - y & = 2, \\ x + 2y & = 1. \end{cases}$$

Exercise 5.14

Solve the system of linear equations and geometrically represent its solution:

$$\begin{cases} x - 2y & = 1, \\ x + 2y & = -1. \end{cases}$$

Exercise 5.15

Geometrically represent this system of linear equations:

$$\begin{cases} x - 2y & = 1, \\ x + 2y & = 1. \end{cases}$$

Exercise 5.16

What are the possible outcomes of a system of linear equations?

Exercise 5.17

Find the value of k so that the line containing the points $(-6, 0)$ and $(k, -5)$ is parallel to the line containing the points $(4, 3)$ and $(1, 7)$.

6. Exercises: Polynomials

Exercise 6.1

Suppose f is a polynomial of degree 55 and its leading term is -1. Describe the long term behavior of this function.

Exercise 6.2

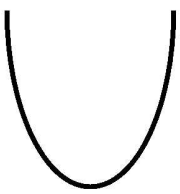
For the polynomial $f(x) = -2x^2(x+2)^2(x^2+1)$, find its x-intercepts.

Exercise 6.3

Find a formula for a polynomial with these roots: 1, 2, and 3.

Exercise 6.4

Is this a parabola?



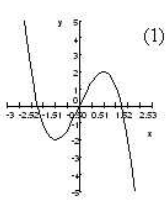
Exercise 6.5

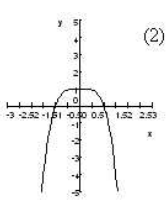
Find the equation satisfied by all points that lie 2 units away from the point $(-1,-2)$ and by no other points.

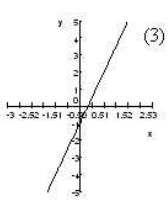
Exercise 6.6

For the polynomials graphed below, find the following:

	1	2	3
smallest possible degree			
sign of the leading coefficient			
degree is odd/even			

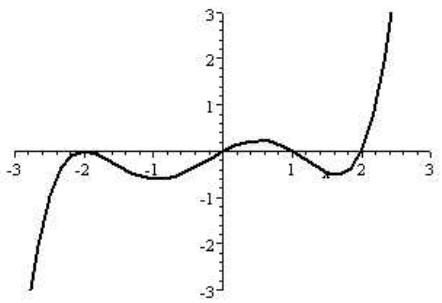






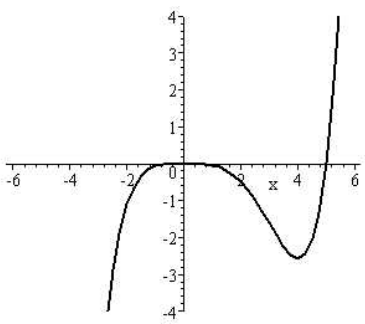
Exercise 6.7

Find a possible formula for the function plotted below:



Exercise 6.8

Find a possible formula for the function plotted below:



Exercise 6.9

For the polynomial $f(x) = -2x(x-2)^2(x+1)^3$, find its x-intercepts and its large scale behavior, i.e., $f(x) \to ?$ as $x \to \pm\infty$.

Exercise 6.10

Given $f(x) = -(x-3)^4(x+1)^3$. Find the leading term and use it to describe the long term behavior of the function.

Exercise 6.11

A factory is to be built on a lot measuring 240 ft by 320 ft. A building code requires that a lawn of uniform width and equal in area to the factory must surround the factory. What must the width of the lawn be?

Exercise 6.12

A factory occupies a lot measuring 240 ft by 320 ft. A building code requires that a lawn of uniform width and equal in area to the factory must surround the factory. What must the width of the lawn be?

Exercise 6.13

(a) Solve the equation $(x^2+1)(x+1)(x-1)=0$.
(b) Solve the inequality $(x^2+1)(x+1)(x-1)>0$.

7. Exercises: Relations and functions

Exercise 7.1

A contractor purchases gravel one cubic yard at a time. A gravel driveway x yards long and 4 yards wide is to be poured to a depth of 1.5 foot. Find a formula for $f(x)$, the number of cubic yards of gravel the contractor buys, assuming that he buys 10 more cubic yards of gravel than are needed.

Exercise 7.2

Visualize the relation:

$$\frac{x^2}{4} + \frac{y^2}{9} = 1\,.$$

Do you see these 4 and 9 on the graph?

Exercise 7.3

Suppose the cost is $f(x)$ dollars for a taxi trip of x miles. Interpret the following stories in terms of f.

1. Monday, I took a taxi to the station 5 miles away.
2. Tuesday, I took a taxi to the station but then realized that I left something at home and had to come back.
3. Wednesday, I took a taxi to the station and I gave my driver a five dollar tip.
4. Thursday, I took a taxi to the station but the driver got lost and drove five extra miles.
5. Friday, I have been taking a taxi to the station all week on credit; I pay what I owe today.

What if there is an extra charge per ride of m dollars?

Exercise 7.4

Let $f : A \to B$ and $g : C \to D$ be two possible functions. For each of the following functions, state whether or not you can compute $f \circ g$:

- $D \subset B$
- $C \subset A$
- $B \subset D$
- $B = C$

Exercise 7.5

An amusement park sells multi-day passes. The function $g(x) = 1/3x$ represent the number of days a pass will work, where x is the amount of money paid, in dollars. Interpret the meaning of $g(6) = 3$.

Exercise 7.6

The perimeter of a rectangle is 10 feet. (a) Express the area of the rectangle in terms of its width. (b) Find the minimal possible area. (c) Find the maximal possible area.

Exercise 7.7

Let $A = f(r)$ be the area of a circle with radius r and $r = h(t)$ be the radius of the circle at time t. Which of the following statements correctly provides a practical interpretation of the composition $f(h(t))$?

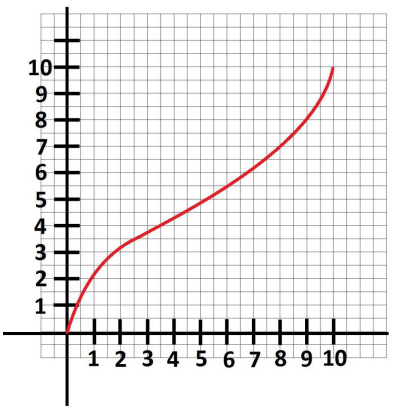
1. The length of the radius at time t.
2. The area of the circle at time t.
3. The length of the radius of a circle with area $A = f(r)$ at time t.
4. The area of the circle which at time t has radius $h(t)$.
5. The time t when the area will be $A = f(r)$.
6. The time t when the radius will be $r = h(t)$.

Exercise 7.8

The area of a rectangle is 100 sq. feet. (a) Express the perimeter of the rectangle in terms of its width. (b) Find the minimal possible perimeter. (c) Find the maximal possible perimeter.

Exercise 7.9

The graph of the function $y = f(x)$ is given below. (a) Find such a y that the point $(2, y)$ belongs to the graph. (b) Find such an x that the point $(x, 3)$ belongs to the graph. (b) Find such an x that the point (x, x) belongs to the graph. Show your drawing.



Exercise 7.10

Make a flowchart and then provide a formula for the function $y = f(x)$ that represents a parking fee for a stay of x hours. It is computed as follows: free for the first hour and \$1 per hour beyond.

Exercise 7.11

Find all possible values of x for which

$$\tan x = 0\,.$$

Exercise 7.12

Make a hand-drawn sketch of the graph of the function:

$$f(x) = \begin{cases} -3 & \text{if } x < 0, \\ x^2 & \text{if } 0 \le x < 1, \\ x & \text{if } x > 1. \end{cases}$$

Exercise 7.13

Find the implied domains of the functions given by:

$$\text{(a) } \frac{x+1}{\sqrt{x^2-1}}; \quad \text{(b) } \sqrt[4]{x+1}\,.$$

Exercise 7.14

Find the implied domain of the function given by:

$$\frac{1}{(x-1)(x^2+1)}\,.$$

Exercise 7.15

Find the implied domain of the function given by:

$$\frac{1}{\sqrt{x+1}}\,.$$

Exercise 7.16

Find the implied domain of the function:

$$\frac{x-1}{x+1}\ln(x^2+1)\sin x\,.$$

Exercise 7.17

Find the implied domain of the function:

$$(x-1)(x^2+1)2^x\,.$$

Exercise 7.18

Finish the sentence: "If a function fails the horizontal line test, then..."

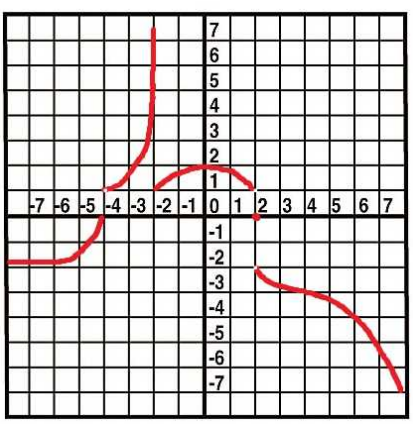
Exercise 7.19

Restate (but do not solve) the following problem algebraically: "What are the dimensions of the rectangle with the smallest possible perimeter and area fixed at 100?"

Exercise 7.20

A sketch of the graph of a function f and its table of values are given below.
Complete the table:

x	0		3		1
y	2	4		5	



Exercise 7.21

Plot the graph of the function $y = f(x)$, where x is the income (in thousands of dollars) and $f(x)$ is the tax bill (in thousands of dollars) for the income of x, which is computed as follows: no tax on the first \$10,000, then 5% for the next \$10,000, and 10% for the rest of the income.

Exercise 7.22

Plot the graph of the function $y = f(x)$, where x is time in hours and $y = f(x)$ is the parking fee over x hours, which is computed as follows: free for the first hour, then \$1 per every full hour for the next 3 hours, and a flat fee of \$5 for anything longer.

Exercise 7.23

Explain the difference between these two functions:

$$\sqrt{\frac{x-1}{x+1}} \text{ and } \frac{\sqrt{x-1}}{\sqrt{x+1}}.$$

Exercise 7.24

Classify these functions:

function	odd	even	onto	one-to-one
$f(x) = 2x - 1$				
$g(x) = -x + 2$				
$h(x) = 3$				

Exercise 7.25

Describe the function that computes the cash-back of 5% followed by the discount of 10%. What if we reverse the order?

Exercise 7.26

Represent this function as a list of instructions:

$$f(x) = \left(\sqrt[3]{\sin x} + 2\right)^{1/2}.$$

Exercise 7.27

Find a formula for the following function:

$$\rightarrow \boxed{\text{square it}} \rightarrow \boxed{\text{take its reciprocal}} \rightarrow$$

Exercise 7.28

Plot the graph of the function given by the list of instructions: 1. add -1; 2. divide by 0; 3. square the outcome.

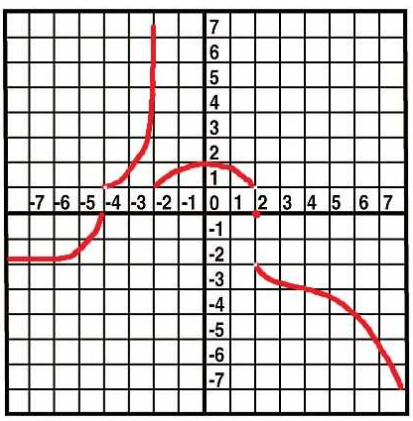
Exercise 7.29

Find the x- and y-intercepts for the graphs in this section.

8. Exercises: Graphs

Exercise 8.1

A sketch of the graph of a function f is given below. Describe its behavior the function using words "decreasing" and "increasing".



Exercise 8.2

Function $y = f(x)$ is given below by a list of some of its values. Make sure the function is onto.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		4	5		2	

Exercise 8.3

Function $y = f(x)$ is given below by a list of some of its values. Add missing values in such a way that the function is one-to-one.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		0	5		0	

Exercise 8.4

What is the relation between being (a) one-to-one or onto and (b) having a mirror symmetry or central symmetry?

Exercise 8.5

By changing its domain or codomain, make the function $y = x^3 - x$ (a) onto, and (b) one-to-one?

Exercise 8.6

Is the function below even, odd, or neither?

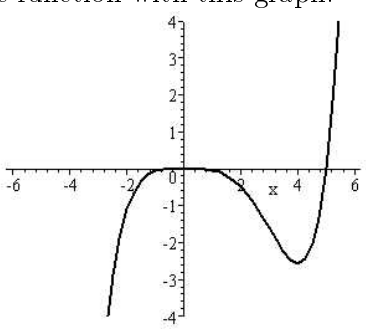
$$f(x) = \frac{x}{e^x - 1} + \frac{1}{2}x - 1$$

Exercise 8.7

Give an example of an even function, an odd function, and a function that's neither. Provide formulas.

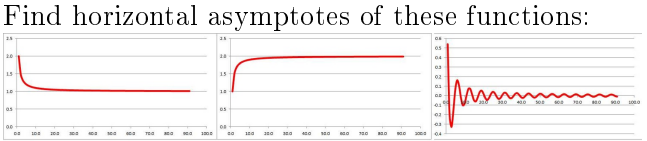
Exercise 8.8

Test whether the following three functions are even, odd, or nether: (a) $f(x) = x^3 + 1$; (b) the function the graph of which is a parabola shifted one unit up; (c) the function with this graph:



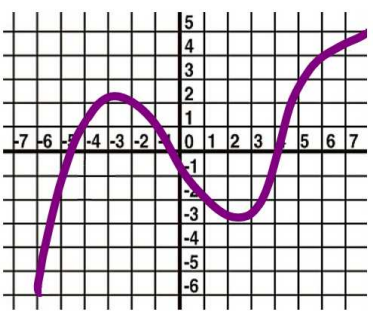
Exercise 8.9

Find horizontal asymptotes of these functions:



Exercise 8.10

The graph of a function $f(x)$ is given below. (a) Find $f(-4)$, $f(0)$, and $f(4)$. (b) Find such an x that $f(x) = 2$. (c) Is the function one-to-one?



Exercise 8.11

Is $\sin x/2$ a periodic function? If it is, find its period. You have to justify your conclusion algebraically.

Exercise 8.12

Is $\sin x + \cos \pi x$ a periodic function? If it is, find its period. You have to justify your conclusion al-

gebraically.

Exercise 8.13

Is $\sin x + \sin 2x$ or $\sin x + \sin \frac{1}{2}x$ a periodic function? If it is, find its period. You have to justify your conclusion algebraically.

Exercise 8.14

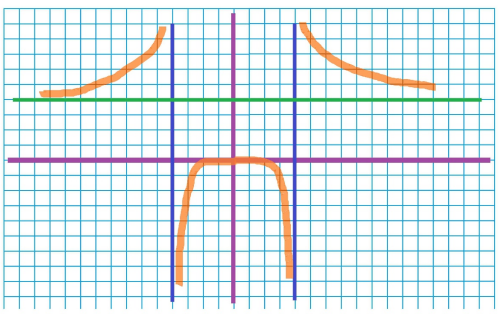
(a) State the definition of a periodic function. (b) Give an example of a periodic polynomial.

Exercise 8.15

Prove, from the definition, that the function $f(x) = x^2 + 1$ is increasing for $x > 0$.

Exercise 8.16

The graph of the function $y = f(x)$ is given below. (a) Find its domain. (b) Determine intervals on which the function is decreasing or increasing. (c) Provide x-coordinates of its relative maxima and minima. (d) Find its asymptotes.



Exercise 8.17

If a rational function has 10 vertical asymptotes, how many branches does its graph have?

Exercise 8.18

For the graph of the function $y = \sqrt{x+8}$, answer the following questions: Is the graph symmetric with respect to the x-axis? The y-axis? The origin?

Exercise 8.19

Determine which of the following statements are true and which are false.

1. The function $\sin x$ on the domain $(-\pi, \pi)$ has at least one input which produces a smallest output value.
2. The function $f(x) = x^3$ with domain $(-3, 3)$ has at least one input which produces a largest output value.
3. The function $f(x) = x^3$ with domain $[-3, 3]$ has at least one input which produces a largest output value.
4. The function $f(x) = x^3$ with domain $[-3, 3]$ has at least one input which produces a smallest output value.
5. The function $\sin x$ on the domain $[-\pi, \pi]$ has at least one input which produces a smallest output value.

Exercise 8.20

Give an example of a function that is both odd and even but not periodic.

Exercise 8.21

Give the definition of a circle.

9. Exercises: Compositions

Exercise 9.1

Represent the function $h(x) = \sin^2 x + \sin^3 x$ as the composition $g \circ f$ of two functions $y = f(x)$ and $z = g(y)$.

Exercise 9.2

Function $y = f(x)$ is given below by a list its values. Find its inverse and represent it by a similar table.

x	0	1	2	3	4
$y = f(x)$	0	1	2	4	3

Exercise 9.3

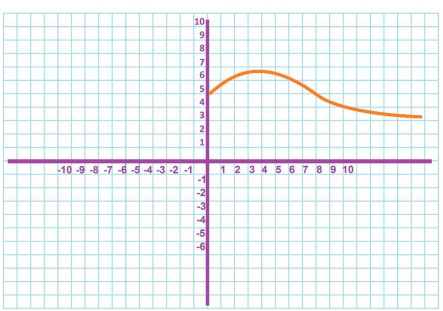
Find the formulas of the inverses of the following functions: (a) $f(x) = (x+1)^3$; (b) $g(x) = \ln(x^3)$.

Exercise 9.4

Are the following functions invertible? 1. $f(n)$ is the number of students in your class whose birthday is on the nth day of the year. 2. $f(t)$ is the total accumulated rainfall in inches t on a given day in a particular location.

Exercise 9.5

The graph of $y = f(x)$ is plotted below. Sketch $y = -f(x+5) - 6$.



Exercise 9.6

Given the tables of values of f, g, find the table of values of $f \circ g$:

x	$y = g(x)$
0	0
1	4
2	3
3	0
4	1

y	$z = f(y)$
0	4
1	4
2	0
3	1
4	2

What if the last rows were missing?

Exercise 9.7

Represent the function below as a composition $f \circ g$ of two functions:

$$h(x) = \sqrt{2x^3 + x}.$$

Exercise 9.8

Find the composition $h(x) = (g \circ f)(x)$ of the functions $y = f(x) = x^2 - 1$ and $g(y) = 3y - 1$. Evaluate $h(1)$.

Exercise 9.9

Represent the function $h(x) = 2\sin^3 x + \sin x + 5$ as the composition of two functions one of which is trigonometric.

Exercise 9.10

(a) Represent the function $h(x) = e^{x^3-1}$, as the composition of two functions f and g. (b) Provide formulas for the two possible compositions of the two functions: "take the logarithm base 2 of" and "take the square root of".

Exercise 9.11

Suppose a function f performs the operation: "take the logarithm base 2 of", and function g performs: "take the square root of". (a) Verbally describe the inverses of f and g. (b) Find the formulas for these four functions. (c) Give them domains and codomains.

Exercise 9.12

1. Represent the function $h(x) = \sqrt{x^2 - 1}$ as the composition of two functions f and g.
2. Provide a formula for the composition $y = f(g(x))$ of $f(u) = u^2 + u$ and $g(x) = 2x - 1$.

Exercise 9.13

Provide a formula for the composition $y = f(g(x))$ of $f(u) = \sin u$ and $g(x) = \sqrt{x}$.

Exercise 9.14

Provide a formula for the composition $y = f(g(x))$ of $f(u) = u^2 - 3u + 2$ and $g(x) = x$.

Exercise 9.15

Find the inverse of the function $f(x) = 3x^2 + 1$. Choose appropriate domains for these two functions.

Exercise 9.16

1. Represent the function $h(x) = \sqrt{x-1}$ as the composition of two functions.
2. Represent the function $k(t) = \sqrt{t^2-1}$ as the composition of three functions.
3. Represent the function $p(t) = \sin\sqrt{t^2-1}$ as the composition of four functions.

Exercise 9.17

(a) What is the composition $f \circ g$ for the functions given by $f(u) = u^2 + u$ and $g(x) = 3$? (a) What is the composition $f \circ g$ for the functions given by $f(u) = 2$ and $g(x) = \sqrt{x}$?

Exercise 9.18

Is the composition of two functions that are odd-/even odd/even?

Exercise 9.19

Represent this function: $h(x) = \dfrac{x^3+1}{x^3-1}$, as the composition of two functions of variables x and y.

Exercise 9.20

Represent the composition of these two functions: $f(x) = \dfrac{1}{x} + 1$ and $g(y) = \sqrt{y-1}$, as a single function h of variable x. Don't simplify.

Exercise 9.21

Function $y = f(x)$ is given below by a list its values. Find its inverse and represent it by a similar table.

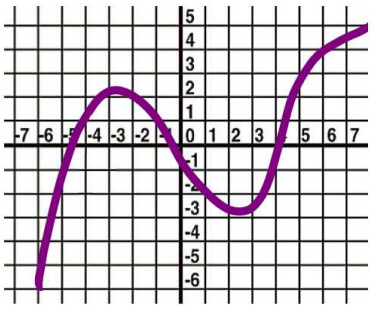
x	0	1	2	3	4
$y = f(x)$	1	2	0	4	3

Exercise 9.22

Give examples of functions that are their own inverses?

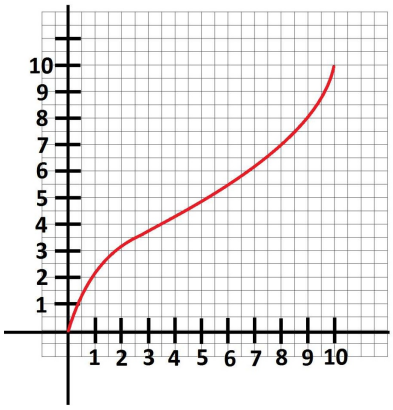
Exercise 9.23

Plot the inverse of the function shown below, if possible.



Exercise 9.24

Plot the graph of the inverse of this function:



Exercise 9.25

Represent this function: $h(x) = \tan(2x)$ as the composition of two functions of variables x and y.

Exercise 9.26

Find the composition $h(x) = (g \circ f)(x)$ of the functions $y = f(x) = x^2 - 1$ and $g(y) = \dfrac{y-1}{y+1}$. Evaluate $h(0)$.

Exercise 9.27

Find the composition $h(x) = (g \circ f)(x)$ of the functions $y = f(x) = x^2 - 1$ and $g(y) = 3y - 1$. Evaluate $h(1)$.

Exercise 9.28

Represent the composition of these two functions: $f(x) = 1/x$ and $g(y) = \dfrac{y}{y^2-3}$, as a single function h of variable x. Don't simplify.

Exercise 9.29

Represent this function: $h(x) = \dfrac{x^3+1}{x^3-1}$, as the composition of two functions of variables x and y.

Exercise 9.30

Function $y = f(x)$ is given below by a list of its values. Is the function one-to one? What about its inverse?

x	0	1	2	3	4
$y = f(x)$	0	1	2	1	2

Exercise 9.31

Function $y = f(x)$ is given below by a list of its values. Is the function one-to one? What about its inverse?

x	0	1	2	3	4
$y = f(x)$	7	5	3	4	6

Exercise 9.32

Functions $y = f(x)$ and $u = g(y)$ are given below by tables of some of their values. Present the composition $u = h(x)$ of these functions by a similar table:

x	0	1	2	3	4
$y = f(x)$	1	1	2	0	2

y	0	1	2	3	4
$u = g(y)$	3	1	2	1	0

Exercise 9.33

Function $y = f(x)$ is given below by a list of some of its values. Add missing values in such a way that the function is one-to one.

x	-1	0	1	2	3	4	5
$y = f(x)$	-1		4	5		2	

Exercise 9.34

Plot the graph of the function $f(x) = \dfrac{1}{x-1}$ and the graph of its inverse. Identify its important features.

Exercise 9.35

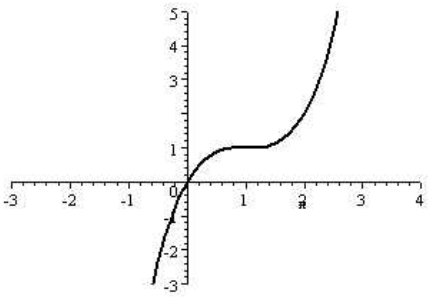
(a) Algebraically, show that the function $f(x) = x^2$ is not one-to-one. (b) Graphically, show that the function $g(x) = 2^{x+1}$ is one-to-one. (c) Find the inverse of g.

Exercise 9.36

Find the formulas of the inverses of the following functions: (a) $f(x) = (x+1)^3$; (b) $g(x) = \ln(x^3)$.

Exercise 9.37

Sketch the graph of the inverse of the function below:



Exercise 9.38

Determine whether the functions below are or are not one-to-one:

$$f(x) = (x-1)^3 \quad \text{and} \quad g(x) = 2^{x-1}\,.$$

Exercise 9.39

Sketch the graph of the composition of the above function and (a) $y = 2x$; (b) $y = x - 1$; and (c) $y = x^2$.

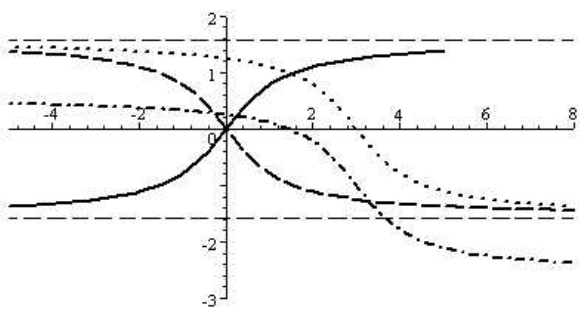
10. Exercises: Transformations

Exercise 10.1

Describe – both geometrically and algebraically – two different transformations that make a 1×1 square into a 2×3 rectangle.

Exercise 10.2

One of the graphs below is that of $y = \arctan x$. What are the others?



Exercise 10.3

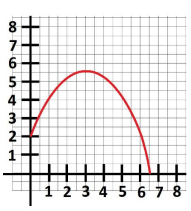
What happens to the domain and the range of a function under the six basic transformations?

Exercise 10.4

How do the six basic transformations affect a function being one-to-one or onto?

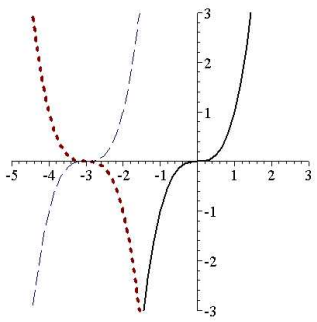
Exercise 10.5

The graph of the function $y = f(x)$ is given below. Sketch the graph of $y = 2f(3x)$ and then $y = f(-x) - 1$.



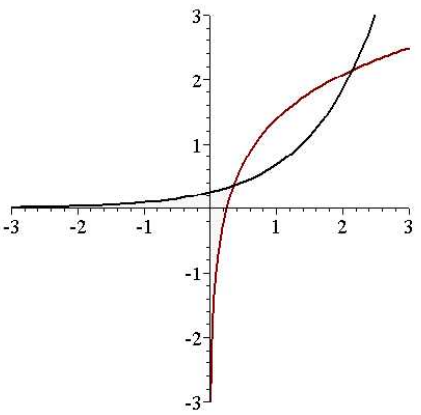
Exercise 10.6

The graph drawn with a solid line is $y = x^3$. What are the other two?



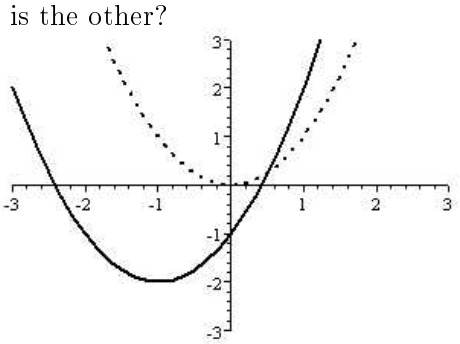
Exercise 10.7

The graph of one of the functions below is $y = e^x$. What is the other?



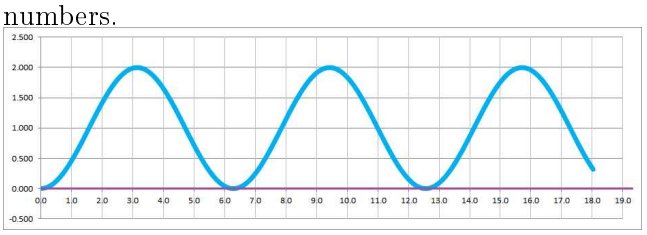
Exercise 10.8

The graphs below are parabolas. One is $y = x^2$. What is the other?



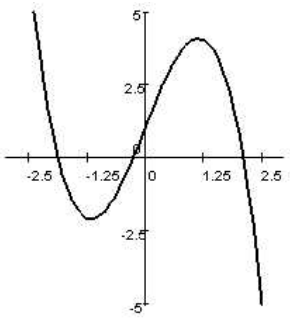
Exercise 10.9

The graph below is the graph of the function $f(x) = A \sin x + B$ for some A and B. Find these numbers.



Exercise 10.10

The graph of function f is given below. Sketch the graph of $y = 2f(x+2)+2$. Explain how you get it.

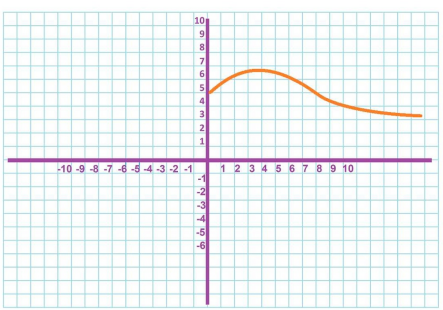


Exercise 10.11

By transforming the graph of $y = e^x$, plot the graph of the function $f(x) = 2e^{x-3}$. Identify the domain, the range, and the asymptotes.

Exercise 10.12

Half of the graph of an even function is shown below; provide the other half:

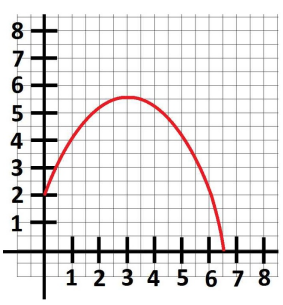


Exercise 10.13

Half of the graph of an odd function is shown above; provide the other half.

Exercise 10.14

The graph of the function $y = f(x)$ is given below. Sketch the graph of $y = 2f(x)$ and then $y = 2f(x) - 1$.

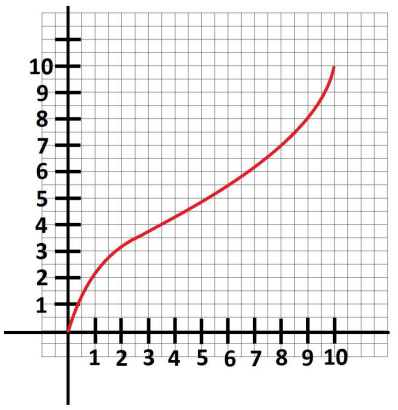


Exercise 10.15

The graph above is a parabola. Find its formula.

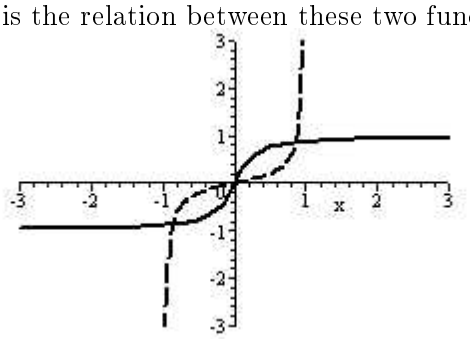
Exercise 10.16

The graph of the function $y = f(x)$ is given below. Sketch the graph of $y = \frac{1}{2}f(x)$ and then $y = \frac{1}{2}f(x-1)$.



Exercise 10.17

What is the relation between these two functions?



Exercise 10.18

Plot the graph of a function that is both odd and even.

Exercise 10.19

Give examples of odd and even functions that aren't polynomials.

Exercise 10.20

Is the inverse of an odd/even function odd/even?

Exercise 10.21

By transforming the graph of $y = \sin x$, plot the graph of the function $f(x) = 2\sin(x-3)$. Identify its range.

Exercise 10.22

Give examples of an even function, an odd function, and a function that's neither. Provide formulas.

11. Exercises: Basic models

Exercise 11.1

The population of a city has doubled in 10 years. Assuming exponential growth, how long does it take to triple?

Exercise 11.2

The population of a city has doubled in 10 years. Assuming exponential growth, how much does it grow every year?

Exercise 11.3

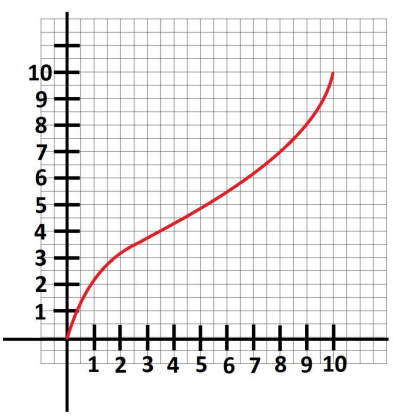
Provide a formula for modeling radioactive decay. What is the half-life of an element?

Exercise 11.4

The population of a city declines by 10% every year. How long will it take to drop to 50% of the current population?

Exercise 11.5

The function $y = f(x)$ shown below represents the location (in miles) of a hiker as a function of time (in hours). Sketch the hiker's velocity as the difference quotient.



Exercise 11.6

A city loses 3% of its population every year. How long will it take to lose 20%?

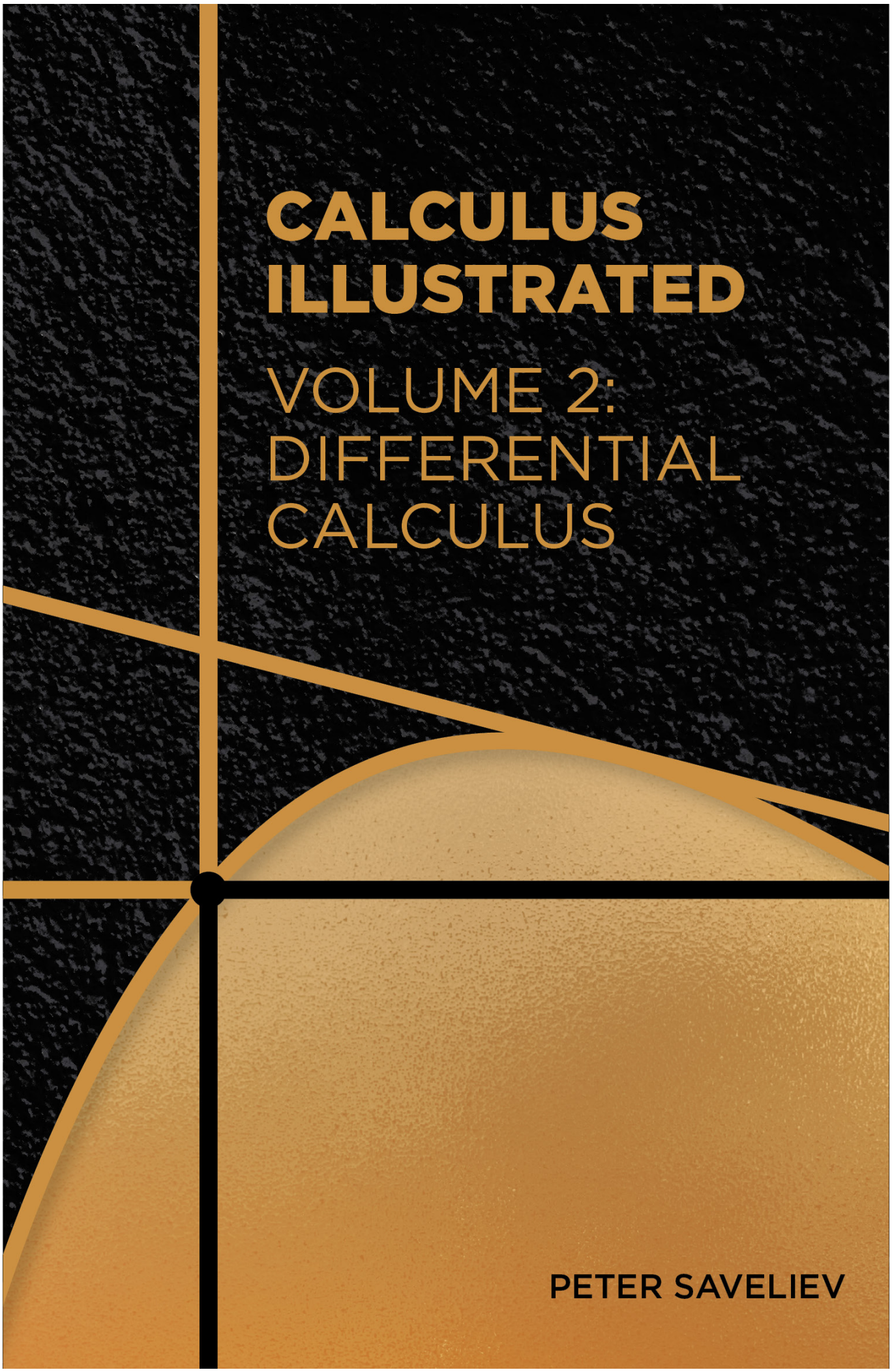
Exercise 11.7

A car start moving east from town A at a constant speed of 60 miles an hour. Town B is located 10 miles south of A. Represent the distance from town B to the car as a function of time.

Exercise 11.8

Sketch the graph of your elevation during a trip on a Ferris wheel.

Next...


CALCULUS
ILLUSTRATED
VOLUME 2:
DIFFERENTIAL
CALCULUS
PETER SAVELIEV

Index

absolute value function, 213
Addition-Multiplication Rule of Exponents, 38, 365
Addition-Multiplication Rule of Logarithms, 419
Additivity for Sums, 84
alternating sequence, 44
amplitude, 460
AND, 204
arccosine, 391
arcsine, 390
arctangent, 392
Arithmetic and Geometric Progressions: Formulas, 37
arithmetic progression, 32, 37, 42, 65

base of exponent, 360
binomial coefficient, 61, 63
Binomial Coefficient Formula, 62
binomial expansion, 58, 60
Binomial Theorem, 60, 76
bound of set, 131
bounded function, 195
bounded set, 131
Boundedness of Trigonometric Functions, 387

Cancellation Laws of Logarithms, 377
Cartesian system, 136, 440, 468, 469, 471, 472, 474
Centered Form of Circle, 449
Choose n From m, 63
circle, 442–444, 449, 450
Circle as Two Graphs, 450
codomain of function, 111
complete square, 286
components of vector, 471
composition, 235, 236, 248, 282, 293–295, 297, 301
Composition of Roots, 356
Composition with Identity Function, 295
Compositions with Constant Function, 293
constant function, 292, 293
constant multiple of function, 399
Constant Multiple Rule for Differences, 94
Constant Multiple Rule for Sums, 95
constant sequence, 28, 65, 68, 80
constant term, 331
Constant Term Is y-Intercept, 332
Continuity of Polynomials, 338
Continuity of Rational Functions, 349
converse, 14
coordinates, 136
cosine, 379, 383, 387, 388, 391, 451, 455–458, 465

decreasing function, 303–305, 308
decreasing sequence, 30
Degree of Factored Polynomial, 334
degree of polynomial, 330, 334
degree of power, 187
degrees, 382
dependent variable, 111
Difference of Arithmetic Progression, 65
Difference of Constant Sequence, 68
difference of functions, 399
Difference of Geometric Progression, 66
difference of sequence, 64–66, 68–70, 92, 94, 96, 97
difference quotient, 183, 190, 367, 389, 476
difference quotient of sequence, 71
Difference under Subtraction, 69
Directions for Dimension 1, 467
Directions for Dimension 2, 472
discriminant of quadratic polynomial, 325, 326
disk, 444
Disk via Inequality, 444
distance, 437, 441, 443, 444
Distance Formula for Dimension 1, 438
Distance Formula for Dimension 2, 440
Distances on Line, 437
Distributive Rule of Exponents, 38, 365
domain of function, 111, 298
Domain of Polynomial, 329
Domain of Tangent, 386
dot product, 472

elements of set, 100
empty set, 126
equal functions, 295, 396
equal sets, 101
equation, 126, 410, 418
Equation of Circle, 443
Equations: Basic Algebra, 410
Equations: General Algebra, 418
equivalence, 14
Even Degree Roots, 355
even function, 316, 388
exponent, 360, 377, 428
Exponent and Logarithm Base Conversion Formulas, 426
Exponent Equal to -1, 53
exponential decay, 427
exponential function, 361, 365, 366, 368, 369, 372, 373, 427
Exponential Function Is Invertible, 369
Exponential Function: Basic Facts, 368
exponential growth, 427
exponential model, 427
extremum of function, 309

Factored Form of Quadratic Polynomial, 325
factorial, 34, 36
factorial of zero, 42

factoring integers, 51
factorization, 334
factors, 305, 334, 338, 349
Formula for Alternating Sequence, 44
Formula for Arithmetic Progression, 42
Formula for Geometric Progression, 42
Formulas of Transformations of Plane, 270
frequency, 461
function, 111, 113, 143, 157, 393, 396
Fundamental Theorem of Algebra, 334
Fundamental Theorem of Arithmetic, 51, 334
Fundamental Theorem of Calculus of Sequences I, 88
Fundamental Theorem of Calculus of Sequences II, 88

geometric growth and decay, 33
geometric progression, 33, 37, 42, 55, 66, 78
global extreme points, 309
graph of function, 154, 165, 450
Graph of Linear Relation, 140
graph of relation, 139, 450

Horizontal Flip, 279
Horizontal Line Test, 157, 230
Horizontal Shift, 276
Horizontal Stretch, 281

identity function, 294, 295
IF-AND-ONLY-IF, 14
IF-THEN, 13
image of function, 193, 232, 328
implication, 13, 14
implied domain of function, 162
inclusion function, 300
increasing function, 303–305, 308
increasing sequence, 30
independent variable, 111
index of terms of sequence, 26
Inequalities: Basic Algebra, 481
Inequalities: General Algebra, 485
inequality, 481, 485
integer value function, 214
intersection of sets, 201, 204
interval, 128, 129
interval notation, 128, 129
inverse function, 245, 246, 248, 297
Inverse of Linear Polynomial, 251
Inverse via Compositions, 248, 297
invertible function, 249, 257
irreducible factor, 326, 334
Irreducible Quadratic Polynomial, 327

leading term of polynomial, 331
Leading Term vs. Degree of Polynomial, 331
Linear Factor Theorem, 335
linear function, 143, 176, 305
Linear Functions, One-to-one Onto, 231
linear polynomial, 176, 251
linear relation, 140, 143
Linear Transformations in Dimension 1, 438
list notation for sets, 100
logarithm, 373, 375, 377, 419, 422, 426, 428

magnitude of vector, 471
maximum of function, 309, 327
maximum of set, 130
Maximum-Minimum of Quadratic Polynomial, 327
minimum of function, 309, 327
minimum of set, 130
monotone function, 303–305, 308, 369
monotone sequence, 30, 68
Monotone vs. One-to-one, 308
monotonicity, 303
Monotonicity and Subtraction, 70
Monotonicity of Exponent, 369
Monotonicity of Geometric Progression, 55
Monotonicity of Linear Functions, 305
Monotonicity of Logarithm, 375
Monotonicity of Sum, 80
Monotonicity Theorem for Sequences, 68
Multiplication-Exponentiation Rule of Exponents, 39, 366
Multiplication-Exponentiation Rule of Logarithms, 422
multiplicity of factor, 51, 335
Multiplicity Rule for Polynomials, 336
Multiplicity Rule for Rational Functions, 347

natural domain of function, 162
Natural Log and Exp Formulas, 428
natural logarithm, 426
negative exponent, 53
Negative Exponent Rule, 54
not equal functions, 296, 396
nth root, 354–356, 365, 366, 368
number line, 22
Number of Permutations, 36
Number of Terms in Binomial Expansion, 58
Number of x-Intercepts, 335
numerical function, 112

Odd Degree Roots, 355
odd function, 318, 388
Odd-Even Trig Functions, 388
One-to-one and Onto vs. Image, 232
one-to-one function, 226, 230, 232, 246, 257, 308, 369
One-to-one Onto vs. Inverse, 246
ONLY-IF, 14
onto function, 225, 230, 232, 246, 257
OR, 204

parabola, 288
Parallel Lines, Same Slope, 469
Parallel Lines: Basic Facts, 469
Pascal Triangle, 59
period, 314
periodic function, 314, 387
Periodicity of Trigonometric Functions, 387
permutations, 36

phase, 461
piecewise-defined function, 208
Point-Slope Form of Line, 178
polynomial, 328–332, 334–338, 343, 344
positively and negatively oriented segments, 467
power functions, 187, 257, 354
Power Functions, Classification, 257
power sequence, 29
preimage of value, 196
prime number, 51
product of functions, 400
Product of Roots, 355
Product Rule for Differences, 96
Pythagorean Theorem, 439
Pythagorean Theorem of Trigonometry, 451

Quadratic Formula, 324
quadratic polynomial, 287, 323–328
quotient of functions, 401
Quotient Rule for Differences, 97

radians, 382
range of function, 193, 328
Range of Linear Function, 193
Range of Quadratic Polynomial, 328
rational exponent, 365, 366, 368
rational function, 343–345, 347, 349
Rational Function's x-Intercepts, 345
reciprocal exponent, 364
recursive, 32–34, 42, 48, 87, 436
relation, 104, 111, 143, 157
repeated addition, 37–39, 41, 51, 54, 334, 344, 360, 368
repeated multiplication, 37–39, 41, 51, 54, 334, 344, 360, 368
Restricted Cosine, 391
Restricted Sine, 390
Restricted Tangent, 392
restriction, 301
restriction of function, 298
Restriction via Compositions, 301
roots of polynomial, 335, 345
roots of quadratic polynomial, 323–326
Rules of Exponents, 372
Rules of Factoring Integers, 51

sequence, 26
sequence as function, 112
sequence of differences, 64, 88, 92, 94, 96, 97
sequence of sums, 73, 76, 88, 93, 95
set, 101
set-building notation, 126, 127, 137, 140, 142, 162, 165, 202, 442
sigma notation, 74
sign function, 212
sine, 379, 383, 387, 388, 390, 451, 455–458, 460, 465
Sine and Cosine of Average Angle, 458
Sine and Cosine of Complementary Angles, 455
Sine and Cosine of Difference, 457
Sine and Cosine of Direction, 468
Sine and Cosine of Double Angle, 457
Sine and Cosine of Half-Angle, 457
Sine and Cosine of Sum, 456
sinusoid, 459
slope, 172, 174, 178, 305, 468, 469
Slope Backwards, 174
Slope From Two Points, 174
Slope is Tangent, 468
slope-intercept form of line, 176
Slopes of Perpendicular Lines, 474
standard form of polynomial, 331
strictly decreasing function, 304
strictly increasing function, 304
strictly monotone functions, 304
subset, 101, 298
substitution, 236
Subtracting Sums of Sequences, 81
Subtracting Sums: Monotonicity, 82
Sum Is Constant Sequence, 80
Sum of Arithmetic Progression, 77
Sum of Binomial Coefficients, 61
sum of functions, 397
Sum of Geometric Progression, 78
sum of sequence, 93, 95
Sum of Sines and Cosines, 458
Sum Rule for Differences, 92
Sum Rule for Sums, 93

tangent, 379, 386, 387, 392, 468
terms of polynomial, 331
transformations, 270, 275, 276, 278–282, 289, 438
Transformations as Compositions, 282
Transformations of Graphs, 289
Triangle Inequality, 441
trigonometric functions, 379, 383, 386–388, 460, 461

unbounded function, 195
union of sets, 200, 204
Uniqueness of Inverse, 246
Unit Circle Equation, 447

variables, 150
vector in dimension 1, 465
vector in dimension 2 , 471
Vertex Form of Quadratic Polynomial, 287
vertex of parabola, 327
vertical asymptote, 349
Vertical Flip, 278
Vertical Line Test, 157
Vertical Shift, 275
Vertical Stretch, 280
Vieta's Formulas, 324

wave function, 460, 461
When Linear Relation Is Function, 143
When Relation Is Function, 111

x-intercept, 170, 338, 349
x-Intercepts of Parabola, 326

y-intercept, 170

Zero Exponent Rule, 40

Zero Factor Property, 334

zero power, 40